Roma Tre University
Ph.D. in Computer Science and Engineering

AI-Based Systems to Help Teachers in Building, Sharing and Sequencing Didactic Materials

Carlo De Medio
Ciclo XXXI

Candidate: Carlo De Medio _____

Advisor: Prof.ssa Carla Limongelli _____

Coordinator: Prof. Stefano Panizieri _____

AI-Based Systems to Help Teachers in Building, Sharing and Sequencing Didactic Materials

A thesis presented by
Carlo De Medio
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Computer Science and Engineering

Roma Tre University
Department of Engineering

15/11/2018

COMMITTEE:
Prof.ssa Carla Limongelli
Prof. Stefano Panzieri

REVIEWERS:
Prof. Damiano Distante, Associate Professor, University of Rome Unitelma Sapienza
Prof.ssa Zuzana Kubincova Associate Professor, Comenius University of Bratislava, Slovakia

Acknowledgments

The achievement of the research doctorate has been a long and difficult path; without the guidance of the many people who followed and helped me I would not be able to complete this great result. I would like to begin by thanking the teacher who followed me from the first day in this faculty, Carla Limongelli who helped me to approach all tasks of university life. I was lucky to meet her as a teacher and as a person and it is the main reason that led me to continue my studies in the Artificial Intelligence Laboratory of Roma Tre. Moreover I owe thanks to all the team with whom I worked in these years composed by Professors Marco Temperini, Filippo Sciarrone and Fabio Gasparetti. I also want to thank my doctoral colleagues with whom I spent three wonderful years of my life between projects and moments of relaxation, Matteo Amedei and Eba. Thanks to my partner Laura and to my whole family that support me in the great difficulties encountered. I also want to thank all the wonderful people known in travels and collaborations, which have make me feel at home in every part of the world. In conclusion I can say that this experience would have been empty if it had not been for all these people with whom I have spent beautiful years and with whom I hope to continue working in the next, and even if it will not be happen I always know I can find in them fantastic friends.

Contents

Chapter 1

Introduction

1.1 Background and Motivations

Thanks to the exponential growth of the Internet, the availability of didactic materials stored in structured form such as in *ad-hoc* repositories or in unstructured forms like HTML pages is strongly increasing. The Internet is becoming a huge repository of materials that can be used by both teachers and learners in order to improve their professional skills. Search engines like Google, Bing or Wikipedia are the most used tools by teachers to find educational materials on the Web [MMK+13]. It is necessary to find and recommend teaching materials to consider all stages of learning [VMO+12]; it is not possible to analyze only the preferences of a user but it becomes necessary to understand the teaching context. Moreover, the use of e-learning in schools and companies has grown exponentially in the last years as reported in [KRDK13]. Consequently, the need to create courses and to retrieve didactic material from the Internet is greatly increased. However, the quality of the teaching material is of fundamental importance. Educational materials are available on Internet in various repositories and in different forms. Over the years, many sites have tried to define guidelines, without defining a standard *de facto*, and searching for the appropriate learning material is a critical issue.

One standing problem in the area of web-based e-learning is how to support instructional designers capability to retrieve, select, and deliver effectively and efficiently learning materials appropriate for their educational purposes, by also speeding considerably up the overall course building process.

Only few platforms offer metadata-based customization but they remain linked to the subjectivity of the teachers. The composition of a course for e-learning platforms,

through the selection and the sequencing of learning materials is a complicated process and is made manually by teachers. In online repositories, correlation between materials and reuse of the knowledge (based on a meta-classifications tagged manually by creator in different formats) is almost impossible. In addition, the Learning Object (LO) scores given by the repositories are non-homogeneous and cause ambiguity. The web is also very rich in educational materials that are not necessarily metadated and that could be used.

The aim of this research is to provide teachers and students with an environment able to search for didactic materials, and their possible relations, without referring to metadata.

One of the most important problems is the organization of information that in the knowledge economy has become a business process of primary importance in many enterprise environments. A well formed information makes easy and quick the retrieving process. Ontology [CJB99] is a versatile technology for organizing information, however, the main obstacle to its full adoption lies in the difficulty of management, especially in its construction and maintenance. In order to overcome this problem it is necessary to resort to the adoption of structured knowledge sources.

Wikipedia is an undisputed source of information for all, especially for students and teachers, although many criticisms have been raised against this type of *unsupervised* information. In 2005, an in-depth study compared Wikipedia with the British encyclopedia, showing the same accuracy percentage [Gil05]. More recent studies distinguish accuracy from completeness, showing that Wikipedia, for its continuous updating, is a more complete source of information than the British encyclopedia and its range of completeness varies between 68% and 91%, reaching low percentages of incompleteness in the pharmacological sector. What is certain is that this huge repository of information cannot always be taken as an a priori truth and it could be risky for students who think they can do self-learning by surfing the pages of Wikipedia.

1.2 Research Goal and Research Questions

Literature study and problem analysis led to three main research questions, which were addressed separately in the following chapters:

- **RQ1:** ***Is it possible to overcome the metadata for retrieving adequate educational material from the web?***
  We have implemented a platform for the creation of courses based on personalized user models. With a series of runtime computations, it generates the knowledge graphs representing the interconnections of the Wikipedia pages associated to the searched concepts, and recommends a subset of those concepts.

The System takes advantage of the classic metrics in the literature such as TF-IDF, IG, LSI and a new metric based on the distance between teaching styles to recommend modules based on user query. The platform is called Wiki Course Builder (WCB [GLS15]) and allows teachers to:

- Create topics and courses using a graphical interface and using an innovative recommendations system,
- Export courses in multiple formats (e.g. pdf, txt, ...),
- Compare through the community graph with other teachers and improve their courses,
- Obtain a Map of the relationships between the chosen materials to optimize the students learning process.

We have performed final experimentation of the system. The channels used to collect the teachers are presented, as well as the structure of the task assigned to each of them. For the future it will be interesting to create a graph, based on the Wikipedia connection graph, with edges that represent the prerequisite/successors relationships. This graph could be a good starting point for a semantic analysis.

- **RQ2:** ***Is it possible to automatically recognize the predecessor/successor relationship between two Learning Object?***
  A first step to answer of this question was starting by exploiting Wikipedia. It is a huge repository of information with high educational content. The study of the relation between Wikipedia pages has highlighted a set of features used for the first test in a classifier trained to recognize the predecessor/successor relationship between two pages. All the metrics used, the feature selection process and the machine learning algorithms applied to find the optimal solution will be described. We present the comparison, in terms of performance, with respect to the systems present in the literature to highlight the performance of the developed classifier.

- **RQ3:** ***Is it possible to automatically find didactic units based on a teaching model?***
  In order to answer this last research question, an external module has been implemented for the retrieving and recommendation of Learning Objects on the Web.

  The external module developed considers teacher-related aspects, in particular the support to LOs retrieval from standard repositories and their reuse. We present a Moodle plug-in that allows teachers to

(i) make queries for searching of learning material (by listing some keywords of her choosing),

(ii) have back a list of LOs taken by several standard repositories,

(iii) see information about the usage of such LOs in the system, if they have been used by the same teacher or her colleagues.

The information shown for the retrieved LOs is related to the "didactic context" in which a LO is used in the courses available in the system (what other LOs are predecessor, or successor, at variable distance, of the given LOs in such courses). This allows to give recommendations based on the actual usage of the LOs in the whole system. We define a teacher model based on the courses defined by the teacher in the system, and on the ways (didactic contexts) of her use of LOs in such courses. The teacher model allows to define a concept of similarity among the teachers working in the system. It also allows to polarize the recommendation system: besides the possibility to give recommendations based on the use of LOs in all the courses of the system, we can also offer recommendations based on a subset of the courses in the system, that is those courses whose teachers are "similar" to the querying one.

1.3 Methodology

The Learning Object Repositories (LORs) are online library for storing, managing, and sharing learning resources; they structure the material on the basis of a Learning Object Metadata (LOM [RSG08]) representation. All the LOs have a variety of metadata that describe the content. In our system, the recommending process is twofold: LOs are retrieved and sorted using the standard TF-IDF metric, implemented by the MyIsam search engine embedded into the MySQL DBMS; subsequently, the returned LOs array is processed using a technique to produce the ranking of items, representing the recommendations to the teachers.
Two major approaches exist in information retrieval (IR): content-based filtering and collaborative filtering (widely described in [LDGS$^+$13], and references therein). A content-based filtering system selects items based on the correlation between the content of the items and the user's preferences, while a collaborative system chooses items based on the correlation between people with similar preferences. When the delivered information comes in the form of a suggestion, an information-filtering system is called a recomender system. We follow the criteria based on a collaborative approach that gives priority to LOs already chosen by users in the community, that are supposed to be more relevant than other retrieved LOs. In fact, in our case, the

community of teachers (e.g. a school or university) shares common background and teaching methodologies.

Also from classical Information Retrieval approaches (TF-IDF, LSI and IG) the corpus is the basic element used by the various processors for calculating the scores. TF-IDF, LSI and IG processors implement the statistical sorting algorithms described in the next chapter, while the new metric (teaching styles) supports the calculation of the Euclidean distance between the teachers teaching styles and the ones associated to the recovered items.

Moreover our purpose is to introduce a content-based approach for identifying prerequisites between text-based LOs (or units of learning materials). This approach makes use of conceptual models represented by weak ontologies. A feature selection methodology allows us to consider the most relevant attributes extracted from the LO content for the topic under consideration. Machine Learning (ML) techniques are considered for recognizing the existence of prerequisite relationships by casting the problem to a binary statistical classification task. Non-formal and implicit classes, categories and relationships that characterize the weak ontology are then exploited in order to infer the requested prerequisite relationship.

Chapter 2

State of the art

This chapter will present studies on information retrieval, sequencing of learning materials and the personalization of online courses that have driven this research. One of the most difficult tasks is the semantic analysis of the teaching materials in order to optimize the recommendation and retrieving processes. Many studies have been done on this argument and will be analyzed in this chapter as well as strengths and weaknesses of the current systems to highlight how the proposed study will overcome the problems using innovative techniques. Initially it will be presented the basic resource of online courses, the learning object, and its uses in classical literature; the methodologies will be discussed, including metadation and sequencing process, to explain how to overcome this structure in online courses and propose an alternative solution. Next, it will be carried out studies concerning the automatic annotation of these materials through algorithms for the automatic extraction of metadata from the text of the materials. Finally, techniques for the retrieving of materials and their classification through ranking algorithms will be presented; those approaches drive the research outside the world of learning objects, passing through a verified and structured knowledge base such as Wikipedia.

2.1 Learning Objects

LOs are the heart of classical online learning systems. The teaching based on the Object Oriented paradigm supports the sharing and reuse of the materials used to build the courses.

The idea came from a suggestion by Reigeluth and Nelson (1997). They noticed that teachers each time they approached materials for the first time, had to break them

down and reassemble them to adapt to their teaching style. This suggested the idea of having a learning unit for each concept so the teachers could share them without having to work on it.

Generally for LO we mean digital entities that can be used via the Internet through an LMS, which allows accessibility and usability to multiple users, whether they are developers or users of the courses. This allows to collaborate in the implementation of the repositories from different places and, using the evaluations of the courses in progress, to trigger a continuous design loop-back to continuously improve the quality of the training provided.

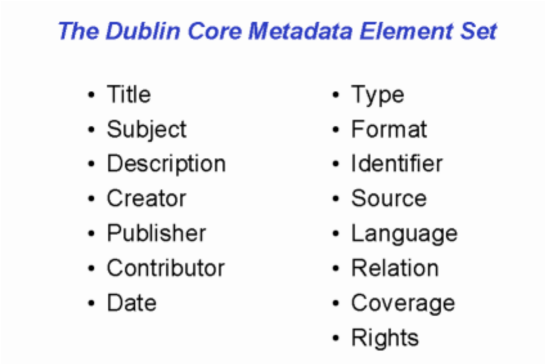
In the Web the use of LOs is strictly linked to the concept of metadata; over the years many standards have been proposed for how to represent metadata but none has been taken as a single model.

The idea of the metadata is to provide a long list of information related to the LO, useful for its classification described by hand by teachers. This information allows an easier retrieving and sharing of materials, fully embracing the Learning Objects paradigm.

Although many models have been proposed to implement metadata, it has not come to a *de facto* standard. There are four most used models:

- Dublin Core Metadata Element Set (ISO Standard 15836): developed in 1995, is a model based on 15 elements described in Fig 2.1. The simplicity of the model allows you to map many of the metadata associated with the online repository materials easily.

Figure 2.1: Dublin Core Elements

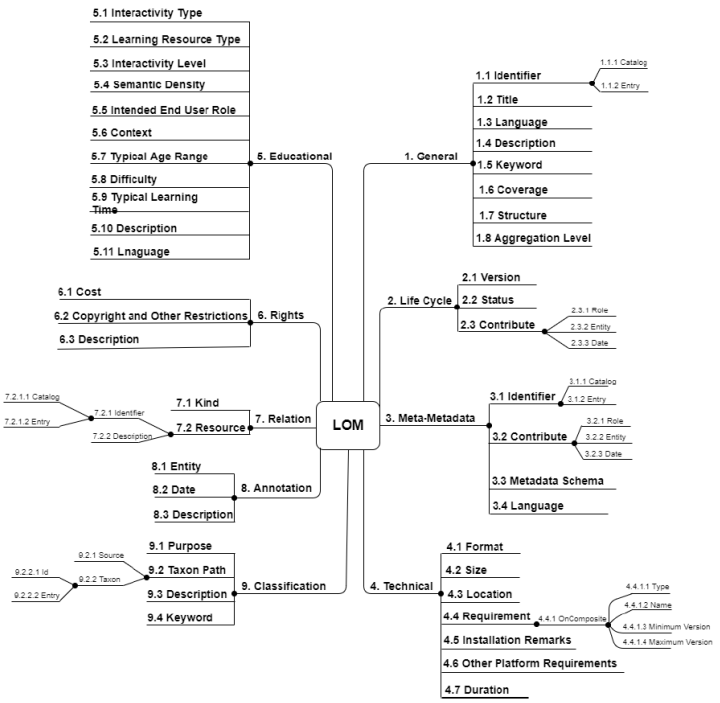


- ISO 19115: 2003 - Geographic Information: Metadata: presented by the geospatial community, is a specific model to represent everything that has a spatial dimension. Describes 14 metadata packages, each with specific functions. Methodologies to extend the model with implementation examples are also discussed.

- PREMIS: Data Dictionary for Metadata Preservation: Model developed on the basis of the Object paradigm in 2005 in accordance with the ISO 14721 OAIS standard; lists the 5 main categories of objects (objects) and the possible relationships between them.

- IEEE LOM see Fig:2.2

Figure 2.2: LOM Schema



2.2 Teaching Styles

The teaching style identifies the general principles, pedagogical and management strategies used for the education of a class of students. A first distinction is made between two macro-categories:

- the teacher-centered approach,
- the learner-centered approach.

In the first case, the teachers constitute the main figure, where the students are seen as "empty-containers", whose main role is to passively receive information (through lessons or direct instruction) with the final goal of overcoming tests or exams. In this model, teaching and examination are seen as two separate entities; student learning is measured through test and exam scores. Instead, the student-centered approach is based on an active and equal role of teachers and students in the learning process. The primary role of the teacher is to guide and facilitate the student's learning through the overall understanding of the teaching material. Students learning is assessed with both formal and informal tests, which also include group projects and collective participation of the class. Teaching and exercise are connected: the student is continuously measured during the teaching process. In order to better understand both approaches, it is important to analyze the three main teaching / learning styles recognized in pedagogy: direct education, research based learning and cooperative learning. The 5 teaching figures theorized by Grasha in 1996 [A.G96] are presented in the this chapter: Expert, Formal Authority, Personal Model, Facilitator and Delegator.

Direct education is the general term used to identify the traditional teaching strategy, which is based on explicit teaching through lessons and demonstrations conducted by the teacher. Direct education is the primary teaching strategy in the teacher-centered approach, in which professors are the only providers of knowledge and information. This teaching method is strongly reflected in the Formal Authority, Expert, and Personal Model styles. Research-based learning is a teaching method that focuses on the student's investigative skills and active participation in learning. In this method, the main role of the teacher is to be a facilitator, ensuring guidance and support to the students during the learning process. This method falls within the learner-centered approach, in which the student has an active and participatory role in his own learning process (for example the webquest method). In this case, it is the styles Facilitator, Personal Model and Delegator to have a greater contribution [A.G96]. Cooperative learning refers to a method of teaching and managing the class that emphasizes teamwork and encourages a strong sense of community. This model fosters students' social and academic growth and includes techniques such as Think-Pair-Share (a cooperative

discussion strategy, developed by Frank Lyman) and mutual teaching. Co-operative learning also falls within the learner-centered approach, as students are given responsibility for their education and development. This method focuses on the belief that students learn better when they work with peers and learn from them. The styles Facilitator and Delegator are part of this category.

The main characteristics of the five figures defined by Grasha are reported:

- Formal Authority: The teachers with the predominant Formal Authority style exploit their position of power and authority, guaranteed by their superior knowledge and status. The style of management of the class is traditional and based on rules and expectations established by the teacher, which provides students with a learning structure. The advantage of this approach is to focus on clear expectations and acceptable methods. On the other hand, the excess of authority could lead to a strong management of the students and their problems.

- Expert: The teachers belonging to the Expert category are in possession of all the knowledge and experience necessary for their students. Their primary role is to guide and direct learners throughout the learning process. Students are only seen as the receptors of knowledge and information ("empty containers"). The professor, as an expert, challenges students to improve their skills, concentrates on transmitting information and requires students to be predisposed to learning and subsequently using information. The advantage of this approach lies in the combination of the teacher's abilities, skills and knowledge, which are transferred to the learners, while the main disadvantage consists in the fact that excessive ostentation of knowledge could intimidate the less experienced students.

- Personal Model: In this case, the teachers conduct the lessons through examples, demonstrating to students how to better access and understand the information. Through this style of teaching, students learn thanks to the observation and the emulation of the process performed by the teacher; he establishes prototypes of thoughts and behaviors, that the students must emulate. The advantage of this style is the emphasis given to the direct observation and emulation of a model's rules. However, some teachers might consider their approach to be "the best way", leading some students to feel inadequate if they fail to meet the expectations and standards of the methods they observe.

- Facilitator: The facilitator teacher gives great importance to the teacher-student relationship. Operating according to an open class model, the teacher's instructions are less important. The cognitive process of the learner is guided in a

gentle way by the educator and is focused on the promotion of independence, on learning by direct experience and on exploration. The advantage lies in the fact that the teacher focuses on the real needs and objectives of the student, allowing them to explore the different options with alternative methods. The main disadvantage comes from the fact that it is a time-consuming method.

- Delegator: This professor develops the ability of students to work independently, encouraging them to work on their projects as part of a team. The teacher, however, remains available on explicit request, playing a passive role in the learning process. The main objective is to promote the sense of autonomy in the educational process; with the obvious advantage of helping students perceive themselves as independent learners. On the other hand, it could lead the teacher to misunderstand the availability of the student towards independent work. Some students may become anxious when they are given too much autonomy.

2.3 Learning Styles

Teaching styles outline the figure of the teacher and what types of educational pathways he prefers. In addition to these, learning styles have been studied, capable of profiling preferences in the type of material that students are more inclined to learn. In this context there are three models known in literature.

- Kolb learning [KK05]. Kolb's model is based on learning from experience. In this model the learning process is cyclical and consists of four phases:

  (i) concrete experience: the user makes a direct experience (e.g tutorial, game, ...),

  (ii) reflective observation: in this phase the details on direct experiences are collected,

  (iii) abstract conceptualization: the reflections made in the previous phase are integrated with previous theories and knowledge,

  (iv) active experimentation: the cycle closes with field tests and checks of what has been learned in new situations.

- Howard Gardner [Gar05]: Gardner starts from the belief that the classic theory of intelligence, measurable through IQ is wrong. He identifies seven types of independent faculties:

(i) logical-mathematical intelligence: ability used in the comparison and evaluation of concrete or abstract objects and in identifying relationships and principles,

(ii) linguistic intelligence: ability expressed in the use of language and in the mastery of linguistic terms,

(iii) spatial intelligence: the ability to perceive and represent visual objects, ideally manipulating them even in their absence,

(iv) musical intelligence: the ability to compose music and accurately analyze the height of sounds and rhythms,

(v) kinesthetic intelligence: ability to control movements and body coordination.

(vi) interpersonal intelligence: ability to interpret the emotions, motivations and moods of other people,

(vii) intrapersonal intelligence: ability to understand the emotions and to put them into socially acceptable forms. Moreover, Gardner has subsequently added an eighth kind of intelligence, that is the naturalistic one, concerning the recognition and classification of natural objects.

- Grasha-Riechmann [BTM14]: Grasha and Riechmann define learning as a process of a social nature and influenced by different individual approaches in the classroom environment. Some dichotomies have been considered that can be considered as extremes of a continuum of socio-relational attitudes. The main categories identified are:

  (i) competition vs collaboration: the student's motivation to learn can be competitive (emerge from the group) or linked to the need to collaborate with other students by sharing experiences and knowledge,

  (ii) intrasubjectivity vs. intersubjectivity: categories relating to the perception of self in relation to the environment and to the culture of belonging. In the learning process, an intrasubjective personality prefers self-analysis as opposed to an intersubjective that will highlight the contribution that the group can give to its personal growth,

  (iii) independence vs addiction: the dependent student sees the teacher as an authority and tends to strictly follow his indications as opposed to an independent one who tries to show a desire for autonomy in the learning process.

The goal of the teaching styles is to create a user model that takes account of the differences that the students can have in the learning process. The fundamental task is to identify a series of features to label the materials in macro-categories (e.g. visual, audio, written, interactive, ...) or to identify attitudes. Platforms based on this model must have a large amount of materials; for every concept the users must insert at least one material that can be traced back to one of the categories. Through an elaboration of the distance between the user model and the values associated with the materials, the best materials are selected for each student and in this way totally personalized courses are obtained. The moodle_LS [LSV11], developed by Roma Tre at Engineering Department realizes all these functions within the Moodle LMS; the creation of the user model is entrusted to a questionnaire and the distance between the model and the materials is the Euclidean distance between the vectors.

The limits of this type of systems is the difficulty that teachers have to label materials. Modern Learning Objects are no longer just belonging to a category and it is therefore necessary to find a methodology for labeling them in a objective way. Furthermore, the labeling process is done manually and in order to use these techniques on a large scale it's necessary to have an automatic labeling system.

2.4 Learning Material Annotation

The discovery of Learning Objects from the web is a complicated process and has been the result of several years of study. The greatest difficulty lies in the diversity between the standards described above and different approaches have been developed. Many repositories collect teaching materials and allow search engines to access the desired resources; the most popular are Ariadne,[1], Merlot[2], CNX[3] and Wisk-Online[4]. The next chapters will present the MoodleREC extension. The tool allows to perform cross-searches between repositories using a course-driven user model to improve the recommendations. Another approach is presented by [GGP08] where the authors present different fields for the metadatating process of the materials available on the repositories most used by the teachers and an indexing algorithm to optimize the retrieval. These metadata-based approaches have the limit of not being able to consider all the knowledge present on the web; all articles, knowledge bases, and other resources are not exploitable for the purpose of learning unit retrieving. To overcome

[1] http://www.ariadne-eu.org/
[2] http://www.merlot.org/merlot/index.htm
[3] http://cnx.org/
[4] https://www.wisc-online.com/learn

these limits, many researches are focused on the following two main tasks for the use of all these resources:

(i) retieve and automatically annotate the materials on the web;

(ii) find an optimal sequencing for the learning materials.

The task characterizing the annotation of the Learning Objects is the automatic identification of an encoding that can be managed by the Learning Object calculator. In [MR14] and [MR15] the authors present an approach based on different heuristics built to analyze teaching materials. An inference-rule based parser produces the input for a classification algorithm; this is intended to identify a specific set of metadata as the type of resource (quiz, material, ...) or topic discussed. No semantic analysis is considered in this approach. The results are promising but still linked to a classic metadata approach. Similar is the work of [SFD11] where a precise categorization is defined for all types of Learning Objects such as exercise, definition, course. On this model a tool is presented that is able to exploit these categories to optimize the discovery of the materials. Despite the good results this work wants to try to detach from static categories, therefore it has been examined in the specific mechanism of retrieving. Based on ontologies existing in [DJ13] the authors try to improve the performances of automatic retrieval of materials from online repositories. [RLM07] presents a project for the automatic semantic classification of all educational audio-visual materials. Through Speech recognition and text analysis they get semantic metadata.

A very onerous process is the disambiguation of educational materials from the rest of the web. There are many fake news or materials inserted by non-expert users; many studies have been carried out for the recognition of teaching materials. In [AL07] the problem is addressed through the identification of a series of features called educational indicators considered as the DNA of the teaching materials. A search engine based on this study was also developed named SaxSearch [AL07].

2.5 Recognition of Prerequisite Relationship and Sequencing

In the literature there are different approaches to the problem with unsatisfactory results. An example is presented in the work [SSG14] where the authors use random models to be able to recognize the relationships between concepts of averages of latent values. However, this method does not take into consideration the content of the teaching materials but takes the values of the tests on the students after the courses as measures. Otherwise the authors in [RSG08] define the relationship between concepts learned from a material and the concepts necessary for learning. In the research

it is analyzed how particular parts of a sentence can be indicators of concepts; on this hypothesis a series of characteristic features are described, such as the length of the material and the nouns related to the prerequisites.

One of the most interesting approaches, and the one that initially led the research, is [LWHG15] where the authors try to exploit Wikipedia in order to recognize the prerequisite relationship between Learning Objects. The TF-IDF metric is calculated on the Wikipedia pages associated with the material and a statistic is created to recognize the prerequisites. The fundamental problem of the approach is in the extraction of the annotations through simple rules and the precision of the predictions, that is very low.

Another approach based on the Wikipedia exploitation is [JP13]. In this work the authors show how, by analyzing the graph of connections between Wikipedia pages, it is possible to infer the prerequisite relationship between them. The study is based on the amount of incoming and outgoing links between pairs of pages. The approach was not supported by experimentation, but the results presented in the following chapters partially validate the approach.

The work [WL16] takes up the previous works by creating an ontology for the recognition of the prerequisite relationship based on a set of three fundamental features: number of incoming links, number of outgoing links, average lengths. Unlike these approaches in [YLCM15] the author proposes to use Wikipedia taxonomy to identify the prerequisite relationship between courses; the idea is to create curricula automatically.

2.6 Wikis as Repository

Wikis provide large repositories that allow students to learn by constructing knowledge in a self-directing manner as described in [PM09] [DK12]. However, there is little research on wikis as large learning environments, and on approaches which leverage the possibilities that Wiki systems offer for didactic scenarios. It comes as no surprise since wikis are not designed as learning environments for formal education. Pusey and Meiselwitz discuss assessment practice when wikis are used as learning environments in higher education [PM11]. Johnson and Bartolino [JB09] proved how student-authored wikis are helpful in building community among incoming students. This Wiki interactive pages model of collaboration allows the students to actively work on the same materials online and helps them to grow academically. Reinhold [Rei06] suggests a lightweight augmentation of the basic structure and navigation of wikis by means of trails, or paths, as guides through the content. This information is generated by tracking the navigational behavior of the individual Wiki users to determine the pages that visitors view, as well as the order of such views. Sequencing

methods and techniques have been widely investigated. They can be organized in two categories. The first category represents the approaches that plan the entire learning path at the beginning of a course, modifying it if the study does not succeed as it should. e.g., [BP03] and [ST13]. In the second category the sequencing is obtained in an implicit way, step by step, through adaptive navigation support techniques, such as adaptive link annotation and direct guidance as shown in [Bru01] and [LMST12]. As distinctive feature with respect to the state-of-the- art, the proposed approach devises learning paths on the basis of the teacher model only, without the need to explicitly represent the student model. Many attempts have been proposed to use Wikipedia as a didactic source, to extract useful information about the relevance of its contents. In particular, in [MW13] an interesting toolkit to manage the Wikipedia contents from a semantic point of view is presented. In [SP06] a comparison between Wikipedia and WordNet is presented in the WikiRelate system, to find semantic relationships among terms. In [GM09] a Wikipedia-based semantic interpretation for natural language processing is presented. A novel method, called Explicit Semantic Analysis, for fine-grained semantic interpretation of unrestricted natural language texts, is presented. This method represents meaning in a high-dimensional space of concepts derived from Wikipedia, representing the meaning of any text in terms of Wikipedia-based concepts. Another work worth of mention is Turchi et al. [TMCO15] where the MediaWiki search engine, made available by Wikimedia Foundation to search contents among Wikipedia web pages, is used to test a ranking algorithm based on Swarm Intelligence.

2.7 Sql vs no-sql

The relational database, born in the 70s, are the most used solution in systems from all the world. The main motivations that naturally lead to the use of a relational database can be summarized in the following:

- provides an easy way to store large amounts of data;
- the entities-relation model on which they are based allows a quick retrieval of data;
- often the most complex systems are the result of the cooperation of heterogeneous systems; The SQL language integrated in all programming languages allows interoperability and exchange of data between these systems;

- the SQL language remains decoupled from the database provider (postgress, mysql, etc.) and allows the users to develop their own applications without knowing the particular system on which they works.

Despite of this great success, with the passing of the years, the first problems caused by the new needs of the systems began to arise. The enormous growth of the internet has caused an exponential growth of websites and online platforms. The servers were no longer able to manage the entire load on their own and so the clusters and load balancers were developed to be able to make up for these problems. In this period it was realized that relational databases had not been designed to work with clusters. With the start of the new millennium, the search for new ways to store and organize data began. The first developed products are described in the publications on the Big Table project by Google (Bigtables [CDG+08]) and the Amazon Dynamo ([Siv12]). Initially it was a methodology to save data in a shared way on clusters, but the research came to conceive the first NOSQL datatbase.

In 2009 a meeting was organized in San Francisco where for the first time the NOSQL technologies were presented including mongoDB[5], Cassandra[LM10] and Hypertable[RC12]. All these systems have the following characteristics in common:

- They are mostly Open Source projects.
- The Independence from the relational model allows greater freedom for developers. It's not necessary to define everything but it is only necessary to enter data within the collections.
- The development of these approach has been driven by the problems found on the web, so they are more performing and easy to integrate into modern systems.
- Since an element contains all the necessary information, you do not need to use the expensive (in terms of performance) JOIN as it happens for the relational databases. A comparison of the performances between SQl and NOSQL databases is presented in the work [GGPO15] such as the advantages of using a non-relational database compared to a relational database.

[5] www.mongodb.com

Chapter 3

Prerequisite Relationship Recognizer

Before presenting the research process that led to the development of the prerequisite recognizer some definitions are reported to understand the formulations of the identified problems. Leaning Objects are represented by their textual content T extracted from a module able to recognize the most used formats on the web and save only the text excluding the not relevant parts for evaluation purposes. This module can parse textual items, pdf file or html pages. Given two LO L_1 and L_2, the prerequisite relation is defined as LO_1 -> LO_2 when the concept associated with LO_1 is fundamental to the learning of the concept associated to the LO_2. In the same way we define the succesor relation LO_1 <- LO_2, representing the opposite concept.

The relationships can be analyzed individually because if there exists LO_1 -> LO_2 the opposite relationship can not exist and if neither of the two relationships are recognized, it is possible to state that between the two LOs there is no relation. We therefore define the function

$$g : \mathbb{L} \times \mathbb{L} \rightarrow Y$$

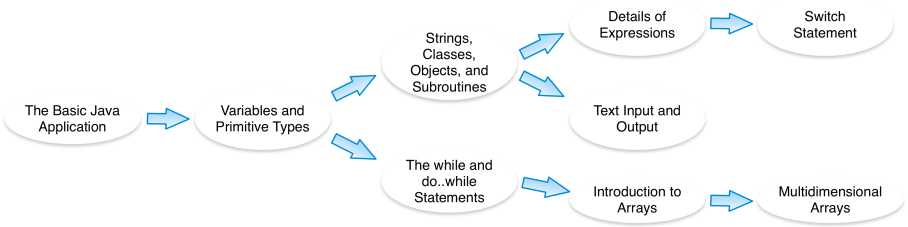
where $\mathbb{L}$ is the set of LOs and Y = {->, 0}; -> indicate the prerequisite relationship between two LOs (i.e., first LO is prerequisite of the second) and 0 that no prerequisite exists.

Fig 3.1 shows an example of prerequisite between Learning Objects.

The problem can be traced back to a classic problem of binary classification, where two elements must be associated with a prediction through a series of determining factors.

The goal of these approaches is to obtain a model that using a training set of correctly classified instances can answer to the problem. The lack of this type of dataset is analyzed in [GMS09] where the authors propose a novel approach for the automatic

Figure 3.1: prerequisite relationships



extraction of relationships from Coursera[1], an online platform with more than 3000 courses, exploiting the hypothesis that the order of materials within the MOOC materials can provide indications on prerequisite relations. The results are promising but for this research it was decided to use datasets labeled by hand by domain experts in order to minimize the noise introduced during the experiments and to provide more accurate evaluations. Moreover exploiting Coursera in [ECLLM16] authors create a knowledge base with relationships between entities in order to generate a semi-taxonomy.

Another type of approach is presented in [SSG14] where the authors try to recognize the prerequisites relationships between materials by analyzing the results of the final quiz of the courses made by the students. The approach, unique in its kind, is interesting, but the results are too much domain dependent and it is not possible to generalize a single model.

3.1 Research Approach

The goal of the first part of the research was the development of a classifier with optimal performance to calculate the function g[3]. The chosen approach was to construct an ontology for the classification of LO by exploiting a comprehensive and trustworthy knowledge base; it was decided to exploit Wikipedia and exploiting the relationships between the pages associated with each LO, to infer the existence of the searched relationships. Specifically, TAGME [2] - an online material annotation service was used to obtain the set of Wikipedia pages associated with a material: $LO = W_1 ... W_k$.

[1] www.couresera.it
[2] https://tagme.d4science.org/tagme/

The first steps was described in [MGLS17][DMGL$^+$16b][DMGL$^+$16a] were the following seven hypotheses for the recognition of relationships were formulated:

(i) H1: The initial sections of the Wikipedia pages associated with the pages have the most information content. The concepts in these sections can be the most characteristic and therefore must have greater weight in the analysis of relationships.

(ii) H2: The relationships between the lengths of two Wiki pages can be indicative of a prerequisite / successor relationship. Through tests we tried to set a threshold between the reports. To calculate the lengths it was decided to exclude all sections that did not contain textual content.

(iii) H3: The analysis of incoming and outgoing links between two sets of Wikipedia pages associated with concepts can highlight a dependence between them and the resulting relationship. This analysis takes into account only the links that from the first set of pages point to the second and are excluded all those external links.

(iv) H4: It is possible to analyze the texts of the links in a set of pages and check the correspondences with the contents of another set. We try to give a stronger definition of dependence compared to that of simple reference. The idea is that if a text of a link appears many times on another page, this can serve to explain it.

(v) H5: The relationship between the number of wiki-pages associated with a LO from the TAGME service is characteristic for relationships. Concepts with many associated pages can be very generic; the more specific concepts, with less pages associated, can be successors of the first ones.

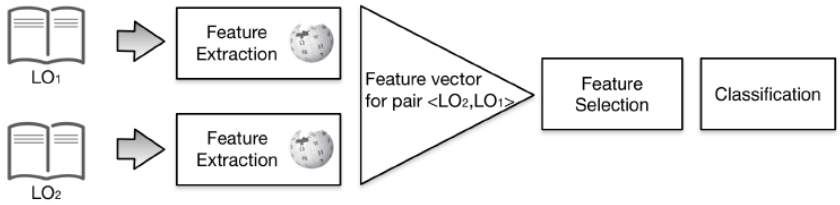
(vi) H6: The study of the text of the materials can suggest the existence of relationships. Specifically, through a logical analysis, strong indicators for classification can be extracted.

(vii) H7: The semantic analysis of two sets of pages can be performed by studying the Wikipedia category associated. It is possible to trace back the categories until you find at least one node in common. The prerequisite / successor relationship can be deduced from the study of the paths on this graph.

3.2 Problem Formulation

For the calculation of the function 3 a supervised learning approach based on two binary classifiers has been implemented [GML+18]. A set of characterizing features described below has been defined. The main steps of the proposed approach are described in Fig 3.2. The characterizing features have been defined and collected in three layers highlighted in Fig.3.3.

Figure 3.2: Principal steps of the proposed approach.



All the features extracted in this process are described in Fig 3.4
The features are then divided into three following categories:

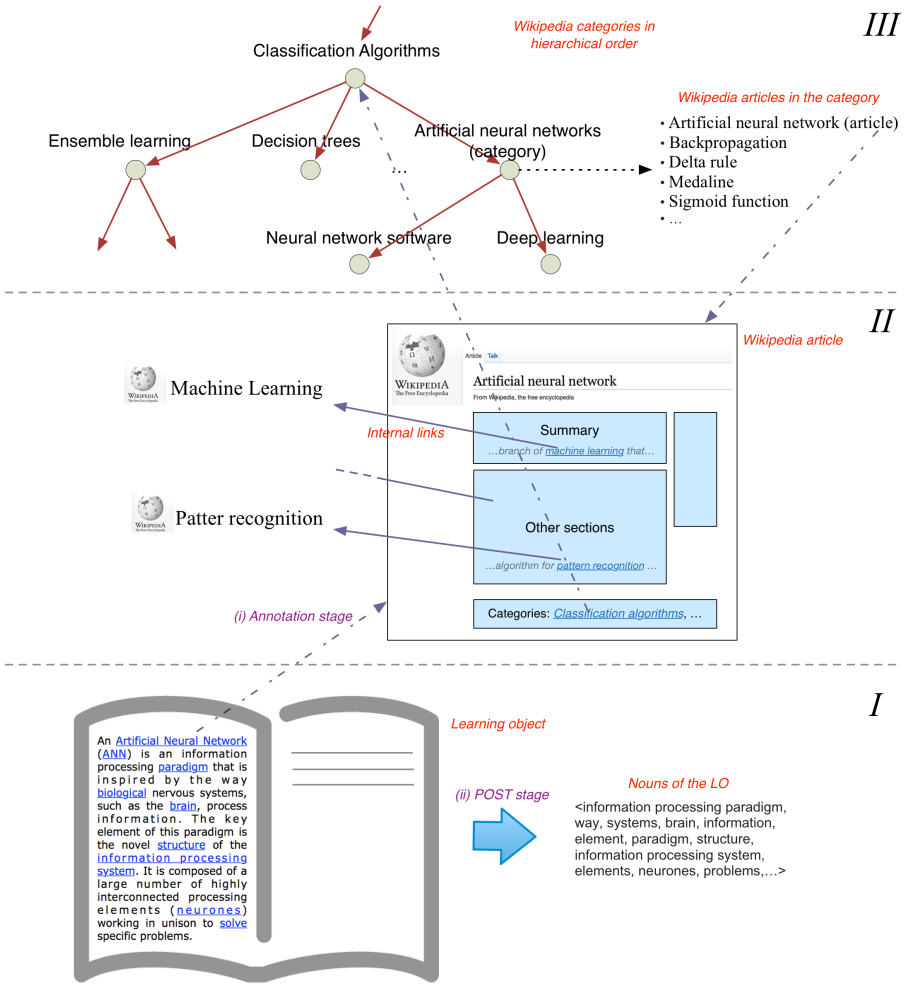
(i) Features extracted from the Learning Object of origin: that is all the measures related only to the material and can be removed from it without further analysis.

(ii) Features relating to a Wikipedia page or a category: all the features in this set are defined on the Wikipedia pages associated with a learning object.

(iii) Features relative to a LO pair: all the measurements relative to the LO pairs are considered. These can not be calculated before runtime but are the most characteristic for the classification. The semantic analysis of the categories belongs to this set.

Starting from the bottom layer a lexical analysis is performed on the text and the basic features such as the length of the page and its sections are extracted. The text is tokenized and saved as a bag-of-word. Word sets are processed by a part-of-speech (POS) tagger that associates a POS to each word (noun, verb, article, adjective, preposition, pronoun, etc). Initially all the elements were saved but from a first analysis it was found that for the purposes of the classification only the nouns gave an information content. In particular, the sets of nouns present in a page are saved and in the

Figure 3.3: The three representation layers of a LO with relevant features



analysis phase the number of the union and intersection sets of the nouns of two pages defined as:

$$f_{nn,\cap}^{(lo_i,lo_j)} = f_{nn}^{(lo_i)} \cap f_{nn}^{(lo_j)} \qquad f_{nn,\cup}^{(lo_i,lo_j)} = f_{nn}^{(lo_i)} \cup f_{nn}^{(lo_j)} \tag{3.1}$$

Figure 3.4: Features definitions

	Features associated with a learning object lo.
$f_{nn}^{(lo)}$	set of nouns in lo.
$f_{C}^{(lo)}$	set of Wikipedia articles annotated to lo.
$f_{l}^{(lo)}$	length in terms of number of words.
	Features associated with a Wikipedia article c *or category* k.
$f_{l}^{(c)}$	length in terms of number of words.
$f_{l'}^{(c)}$	length in terms of number of words of the summary section.
$f_{C,l}^{(lo)}$	average length of the articles in $f_{C}^{(lo)}$.
$f_{C,l'}^{(lo)}$	average length of the summary section of the articles in $f_{C}^{(lo)}$.
$f_{L}^{(c)}$	set of links in c to other articles.
$f_{L'}^{(c)}$	set of links in the summary section of c to other articles.
$f_{t}^{(c)}$	title of the article c.
$f_{K}^{(c)}$	Wikipedia categories assigned to c.
$f_{C,K}^{(k)}$	set of Wikipedia articles in the Wikipedia category k.
	Features associated with a pair $\langle \mathrm{lo}_i, \mathrm{lo}_j \rangle$ *of LOs.*
$f_{L}^{(lo_i,lo_j)}$	set of links in $f_{L}^{(c)}$ that point to $f_{C}^{(lo_j)}$, where $c \in f_{C}^{(lo_i)}$.
$f_{L'}^{(lo_i,lo_j)}$	set of links in $f_{L'}^{(c)}$ that point to $f_{C}^{(lo_j)}$, where $c \in f_{C}^{(lo_i)}$.
$f_{nn,\cap}^{(lo_i,lo_j)}$	set of nouns in $f_{nn}^{(lo_i)}$ that also belong to $f_{nn}^{(lo_j)}$.
$f_{nn,\cup}^{(lo_i,lo_j)}$	set of nouns in $f_{nn}^{(lo_i)}$ and $f_{nn}^{(lo_j)}$.
$f_{C,nn}^{(lo_i,lo_j)}$	number of links in $f_{L}^{(c)}$ whose title corresponds to a POS noun in $f_{nn}^{(lo_j)}$, where $c \in f_{C}^{(lo_i)}$.
$f_{K,d}^{(lo_i,lo_j)}$	counts the number of super-categories or sub-categories that lo_i has in common with lo_j at distance d.

$f_{nn}^{(lo_i)}$ represents the set of nouns extracted from the LO.

A LO is defined as $f_C^{(lo)}$ the set of Wikipedia pages associated by an annotation service. C is the set of wikipedia articles c. Initially the system was based on the service offered by the WikiMedia Fundation, but with the release of the second version of the TAGME service it was decided to migrate to improve the precision of the annotations. In layer II, all the features associated with the Wikipedia pages are calculated. Specifically, the length of the article $f_l^{(c)}$ and its *internal links* $f_L^{(c)}$ are the main ones. In addition, crossed analysis are made between the texts of the incoming links and the nouns of the pages. This measure of correlation is defined as

$$f_{C,nn}^{(lo_i,lo_j)} = \sum_{c \in f_C^{(lo_i)}} \sum_{nn \in f_{nn}^{(lo_j)}} equal(f_t^{(c)}, text(nn)) \tag{3.2}$$

where the *text()* function returns the text of the noun and the *equal()* represents the similarity between two texts. Often in the articles not all the entities present has a links. This analysis identifies these shortcomings and tries to give more weight to the links mentioned many times on another page. This feature links the set of Wikipedia pages associated with the LO with its textual content. The Wikipedia pages are divided into sections and the first is often dedicated to the description of the content. The hypothesis that the content of this section may contain useful information is defined as the notations $f_{l'}^{(c)}$ and $f_{L'}^{(c)}$ as the set of annotations related to the first section or to the full text of the page. All metrics are calculated on both sets and checks have been carried out in order to measure their validity. Measures relating to outgoing links between two pages associated with LOs are defined as

$$f_L^{(lo_i,lo_j)} = \bigcup_{c \in f_C^{(lo_i)}} f_L^{(c)} \cap f_C^{(lo_j)} \qquad f_{L'}^{(lo_i,lo_j)} = \bigcup_{c \in f_C^{(lo_i)}} f_{L'}^{(c)} \cap f_C^{(lo_j)} \tag{3.3}$$

LOs have many articles associated with them by the TAGME service. We defined

$$f_{C,l}^{(lo)} = \frac{1}{\left| f_C^{(lo)} \right|} \sum_{c \in f_C^{(lo)}} f_l^{(c)} \tag{3.4}$$

as the measure constructed calculating the average of the lengths of the associated pages. Excluding the stop words, the tokens deriving from the LO parsed text are counted for this mean. In the same way the measure calculated on the first sections of the Wikipedia pages is defined as

$$f_{C,l'}^{(lo)} = \frac{1}{\left| f_C^{(lo)} \right|} \sum_{c \in f_C^{(lo)}} f_{l'}^{(c)} \tag{3.5}$$

Finally, in the higher layers all the features related to the links between the pages and the relationships existing between them in the knowledge base are extracted. Specifically in the upper layer (see Fig.3.3), a new approach to the semantic analysis of the domains associated to the concepts is presented. The study of the paths among the nodes of the graph of the Wikipedia categories allows to extract a feature representing the generality of a concept with respect to another.

Each article c may be assigned to one or more categories $f_K^{(c)}$, where the k-category contains a set of articles $f_{C,K}^{(k)}$.

Ontologies are defined as links between concepts. Each Wikipedia page belongs to one or more categories within a tree structure created by Wikipedia. For example, Rome, Madrid and London can be classified as european capitals. A taxonomy defines all the relations between its elements. For example the arm entity can be defined as part-of the body entity. The Wikipedia category tree is not organized as a taxonomy but provides a hierarchy of categories. Starting from a root node it is possible to navigate from more general categories upwards towards specific leaves. In some cases between child and parent nodes some relationships can be defined such as *instance-of*, *member-of* or *has-a* if the author described them. In general Wikipedia does not provide information on relationships between entities. We introduce the set of categories K associated by Wikipedia to a page like

$$K^{(lo)} = \bigcup_{c \in f_C^{(lo)}} f_k^{(c)} \tag{3.6}$$

and two functions $childs : \mathbb{K} \to \mathcal{P}(\mathbb{K})$ and $parents : \mathbb{K} \to \mathcal{P}(\mathbb{K})$ that return the two sets containing the top categories in the tree and those below the selected node.

We can compare the upper-category and the down-category between two pairs of lo introducing as a parameter the length of the distance from the category associated to the LO as

$$\begin{aligned}
K_{\uparrow,1}^{(lo)} &= \left\{ k \in \mathbb{K} | childs(k) \in K^{(lo)}) \right\} \\
K_{\downarrow,1}^{(lo)} &= \left\{ k \in \mathbb{K} | parents(k) \in K^{(lo)}) \right\} \\
K_{\uparrow,2}^{(lo)} &= \left\{ k \in \mathbb{K} | childs(k) \in K_{\uparrow,1}^{(lo)}) \right\} \\
K_{\downarrow,2}^{(lo)} &= \left\{ k \in \mathbb{K} | parents(k) \in K_{\downarrow,1}^{(lo)}) \right\} \\
&\dots
\end{aligned}$$

Finally, a single numeric value indicates how many categories above and below the lo_i's categories correspond to the lo_j's categories as follow:

$$f_{K,d}^{(lo_i,lo_j)} = \left|K_{\downarrow,d}^{(lo_i)} \cap K^{(lo_j)}\right| - \left|K_{\uparrow,d}^{(lo_i)} \cap K^{(lo_j)}\right| \tag{3.7}$$

Experimentation has shown that exploiting over distance 2 the Wikipedia category tree is not useful for classification purposes, since the generalization of concepts confuses the predictor as and increasing exponentially the execution time. Positive values of the metric show that one or more categories below lo_i represent the annotations in lo_j.

The hypotheses shown at the beginning of the chapter guided the development of the described metrics; the features can be redundant and introduce noise in the experiments and for this purpose a *feature selection* has been applied using the Information Gain algorithm. Information Gain is the most used method in literature [Mit97] to perform the features selection; analyzes the number of information bits that each feature gives to the prediction, excluding them one at a time. With the expected reduction in entropy in terms of uncertainty. The classical representation of an instance for a classifier is vector and Fig 3.5 shows the representation used for the experiments. For each dimension of the vector has values in $\mathbb{R}$.

Many features presented values in too different ranges. For the generation of instances for the classifier it was decided to normalize these values, especially for the performance of the classifier based on neural network that is established in literature that is more influenced by this type of problems. In the experiments we also considered averages between values calculated separately on two Learning Objects in order to try to identify a correlation between the values.

Figure 3.5: features vector for two LOs

$\frac{\lvert f_C^{(lo_i)}\rvert}{\lvert f_C^{(lo_j)}\rvert}$	$\frac{f_l^{(lo_i)}}{f_l^{(lo_j)}}$	$\frac{f_{C,l}^{(lo_i)}}{f_{C,l}^{(lo_j)}}$	$\frac{f_{C,l'}^{(lo_i)}}{f_{C,l'}^{(lo_j)}}$	$\frac{f_L^{(lo_i,lo_j)}}{f_L^{(lo_j,lo_i)}}$	$\frac{f_{L'}^{(lo_i,lo_j)}}{f_{L'}^{(lo_j,lo_i)}}$	$\frac{\lvert f_{nn}^{(lo_i)}\rvert}{\lvert f_{nn}^{(lo_j)}\rvert}$	$f_{nn,\cap}^{(lo_i,lo_j)}$	$f_{nn,\cup}^{(lo_i,lo_j)}$	$\frac{f_{C,nn}^{(lo_i,lo_j)}}{f_{C,nn}^{(lo_j,lo_i)}}$	$f_{K,1}^{(lo_i,lo_j)}$	$f_{K,2}^{(lo_i,lo_j)}$

3.3 Experimentation

To verify the performances of the classifier, a semi-automatic experiment was organized to measure the accuracy of the predictions. After a period of planning, the following critical issues of the experimentation have been identified:

- The classifier training must be independent of knowledge domains.
- The experiment dataset must be produced by domain experts so as not to introduce noise.
- For the experiment we must find a model in the literature with which to compare in order to verify the effectiveness of the classifier.

To verify the independence from the knowledge domain, the one leave out model was adopted. The whole dataset has been divided into k parts according to the knowledge domains. Specifically, as seen in Fig3.6, each course is seen as a distinct domain. Experimental data are obtained through cross validation: each domain is extracted from the dataset and used as a validation test. All the others are used as a training set and the final results are the arithmetic mean of all the checks done. The main principle is to train a system to recognize reports on general topics and test it in an unknown domain.

The dataset reported in Fig3.6. It is divided into three parts. The first one is taken from the Crowdcomp dataset [LWHG15]: this dataset consists of 5 courses and 1600 Learning Object with 206 specified prerequisite relationships. The dataset has been used for other experiments and provides a textual version of the materials found on the web. The Amazon Mechanical Turk [3] platform was exploited to recruit people used for the manually classification of the instances.

All the teaching materials of the dataset are Wikipedia pages, therefore perfectly suited to work in our framework. In order to validate the classifier, it was decided to take data from other on-line platforms, with materials in different forms to demonstrate the independence from the type of document used. Specifically, courses were taken from two on-line MOOC platforms, among the most widely used in the world, with materials, mainly in video format, coming from University courses:

- EDX [4], with courses from MIT, Harvard, Berkley, University of Texas, Hong Kong Polytecnic University and the University of British Columbia.
- Udacity [5], partner Amazon, Google, IBM and other companies.

To homogenize the dataset, all the materials were translated into simple text files. For video-lessons, official transcripts were requested from Udacity and UDX. Initially a speech recognition module was tested but from the first tests emerged that the error introduced by this module didn't make the results interpretable.

[3] https://www.mturk.com/
[4] https://www.edx.org/
[5] https://eu.udacity.com/

Figure 3.6: Dataset

	ID	*Domain*	*LOs and courses*	*Prerequisites*
CrowdComp	1.	Meiosis	400	67
	2.	Public-key Cryp.	200	27
	3.	Parallel Postulate	200	25
	4.	Newton's Laws	400	44
	5.	Global Warming	400	43
Udacity	6.	Biology	206 (1)	16
	7.	Computer Science	2,396 (4)	68
	8.	Math, Statistics & Data Analysis	1,759 (3)	12
	9.	Physics	546 (1)	10
	10.	Psychology	690 (1)	26
edX	11.	Design	66 (2)	10
	12.	Economy and Finance	91 (2)	12
	13.	Engineering & Project Management	582 (15)	64
	14.	Politics	62 (2)	8

These materials were given to teachers belonging to Roma Tre and Sapienza universities. The task assigned was to physically open the text and report all the couple of materials with a prerequisite relation. 132 classifications were obtained for the Udacity platform and 94 for edX. The courses taken from these platforms also count thousands of materials, so the materials relating to the identified relationships and an equal number of random materials were taken in order to not distort the experiments by putting only instances whose classification is already known. The result of this

process is a dataset containing 432 prerequisite links.

To measure the performance of the classifier, classic metrics proposed by the literature were used:

(i) precision

(ii) recall

(iii) F1 measure

(iv) Area under the ROC curve

The tests were conducted with the Waikato Environment for Knowledge Analysis software [WFHP16]. WEKA is an environment totally written in Java with the most famous machine learning algorithms implemented. The users can create instances and makes the experiments without worrying about the physical implementation of the algorithms. Furthermore, once a trained model is obtained, it is easy to start experiments and save all statistical data.

Three algorithms for the construction of the classifier have been selected, and experiments have been proposed on each one of them so as to verify the most performing:

- Naive Bayes (NB)
- C4.5
- Multilayer Perceptron (MLP)

Furthermore, as a baseline approach we were confronted with a Zero Rule classification (0-RL), which relies on the frequency of targets and predicts the majority target category.

3.4 Evaluation

The results presented in tab 3.1 highlight the usefulness of the proposed approach. All the tested algorithms gave acceptable results, especially the classifier based on Multilayaer Perceptron. The neural network demonstrates that it's able to learn multidimensional maps better than the approaches based on Bayesan classifiers and linear regressions. It is interesting to note how the results of the C4.5 and NB algorithms are complementary: In the first results, higher values of precision and recall are reached, while in the latter we obtain a peak on the AUC curve. The classifier was implemented

Table 3.1: Performance outcomes. Standard deviation σ over the courses inside the parentheses.

	Pr	Re	F1	A	AUC
0-RL	0.34 (0.01)	0.58 (0.01)	0.42 (0.01)	0.68 (0.09)	0.50 (0.00)
C4.5	0.78 (0.01)	0.74 (0.02)	0.74 (0.02)	0.74 (0.02)	0.74 (0.02)
MLP	**0.81** (0.02)	**0.78** (0.03)	**0.78** (0.03)	**0.78** (0.03)	**0.87** (0.01)
NB	0.71 (0.04)	0.70 (0.03)	0.69 (0.03)	0.70 (0.03)	0.78 (0.02)

using the model generated from training with the most performing ML algorithm on all evaluation metrics.

Comparing this approach only with the Crowdcomp dataset, we can see a significant increase of precision compared to 0.61 reached by [LWHG15] (- 28.2%). Extending the experiment to the whole analyzed dataset, it is possible to see how the classifier reaches a precision value equal to 0.81% exceeding the previous approach of 33.1%. This increase is hopeful that the increase of the dataset can further improve the performance. In the comparison approach, the authors associate a set of Wikipedia pages to the analyzed LO pairs; to these pages the algorithm for the calculation of the TF-IDF is applied to recognize the prerequisite relation between the LOs. The approach presented in [LWHG15] is the only other experiment related to my knowledge and the result achieved is incredible. The results of this experiment were published in 2017 in the journal [GML$^+$18]. In order to further improve performance, an additional experiment was carried out. All the instances classified wrong had been collected in a dataset in order to identify characterizing features to reduce the error. An empirical analysis had identified two characteristics common to Learning objects from which the instances derived:

- Many instances are characterized by a very short text, and with many references to previous materials; analyzing the database with the annotations produced by the TAGME service, it was possible to identify a problem in effectively identifying the associated Wikipiedia pages.
- The scientific materials, rich in formulas and graphs, cannot provide characterizing data for the creation of instances for the classifier, in addition the articles in Wikipedia do not contain links useful for predicting the prerequisites / successors.

In order to solve the problems related to the annotations on the teaching materials, two solutions have been proposed; it is possible to generate a dictionary for each domain in order to identify misleading terms in the tokenization process. This can affect

the accuracy of recommendations from the TAGME service. Furthermore, in cases of failed annotation, an expert in the field could be involved for a manual cataloging. The rundown of the specifications of these pages could lead to the identification of new features.

A further analysis was made on the features characterized by the instances for the classifier; two hypotheses were verified:

(i) Is the set of features taken into consideration optimal? Is there a subset of features that can optimize predictions?

(ii) Do some features get better performance within a specific domain?

In 3.7 is possible to see the measure of the Information Gain evaluated on the Croudcomp dataset; most of the information is contained in a subset of the features such as the amount of nouns extracted from the POS Tagger from the Learning Objects and the number of categories in common within the Wikipedia tree. This result is very important for the validation of the semantic analysis on the materials. It is also important to observe that lower values of IG are in correspondence of features related to the single Learning Objects; the most characteristic measures are linked to correlation metrics between materials. The same experiments were conducted, leaving all the features out of the instances one at a time but without obtaining changes in the definitive performances. None of the features presented make predictions worse.

Figure 3.7: Information gain

	$\frac{\lvert f_C^{(lo_i)}\rvert}{\lvert f_C^{(lo_j)}\rvert}$	$\frac{f_l^{(lo_i)}}{f_l^{(lo_j)}}$	$\frac{f_{C,l}^{(lo_i)}}{f_{C,l}^{(lo_j)}}$	$\frac{f_{C,l'}^{(lo_i)}}{f_{C,l'}^{(lo_j)}}$	$\frac{f_L^{(lo_i,lo_j)}}{f_L^{(lo_j,lo_i)}}$	$\frac{f_{L'}^{(lo_i,lo_j)}}{f_{L'}^{(lo_j,lo_i)}}$	$\frac{\lvert f_m^{(lo_i)}\rvert}{\lvert f_m^{(lo_j)}\rvert}$	$f_{m,\cap}^{(lo_i,lo_j)}$	$f_{m,\cup}^{(lo_i,lo_j)}$	$\frac{f_{C,m}^{(lo_i,lo_j)}}{f_{C,m}^{(lo_j,lo_i)}}$	$f_{K,1}^{(lo_i,lo_j)}$	$f_{K,2}^{(lo_i,lo_j)}$
1	0	0	0	0	0	0	0.31	0	0.03	0	0.06	0.01
2	0	0	0	0	0	0	0.17	0	0	0.07	0.04	0
3	0.02	0.08	0	0	0	0	0.16	0.06	0	0	0.25	0
4	0.02	0	0	0	0	0.08	0.32	0	0	0.01	0.12	0
5	0	0.08	0.01	0	0	0	0.20	0	0	0	0.07	0

Finally, Fig.3.8 shows the correlation values of Kendall's τ on all domain pairs d_n, d_m

The results show how for each domain exists a subset of optimal features; The pairs of courses ["Meiosis", "Public-key cryptography"], and ["Global Warming","Newton's

Figure 3.8: Kendalls τ coefficient between each pair of domains d_n and d_m

$$\tau_{d_n,d_m} = \begin{pmatrix} 1 & & & & \\ 0.61 & 1 & & & \\ -0.10 & 0.01 & 1 & & \\ 0.39 & 0.23 & 0.07 & 1 & \\ 0.31 & 0.49 & 0.25 & 0.53 & 1 \end{pmatrix}$$

Laws"] present the most common subset of optimal features. Another very large subset is in common with the couples ["Meiosis", "Parallel Postulate"], ["Parallel Postulate", "Public-key cryptography"] and ["Newton's Laws", "Parallel Postulate "]; the intersection of these two subsets is however formed only by two features that in the previous tests were not the most characterizing on all domains. This proves how the features are characterizing on the basis of the domains and these results with the previous ones show how the set of features identified is optimal.

3.5 Conclusions

A novel approach for discovering prerequisite relationships between text-based learning materials has been proposed. It provides useful support both to instructional designers, in authoring the LO's metadata, and to adaptive learning technologies, which suggest personalized learning paths by sequencing the available learning materials, speeding up the course building operations.

The general-purpose feature-based approach is easily adaptable to various topics by performing standard feature selection techniques. Additional features can also be considered provided that representations of their measures are defined in categorical or numerical form.

Future research activity is towards the identification of semantic relationships and properties associated with Wikipedia resources that can support the identification task. The basic weak relationships provided by the Wikipedia taxonomy are limiting because of their inability to capture domain specific knowledge. Structured databases, such as DBpedia [DBp], allow users to submit semantic queries (e.g., "All the impressionist painters that have actively worked in Netherlands") by sifting through the content spread across many different Wikipedia articles.

Chapter 4

Wiki Course Builder

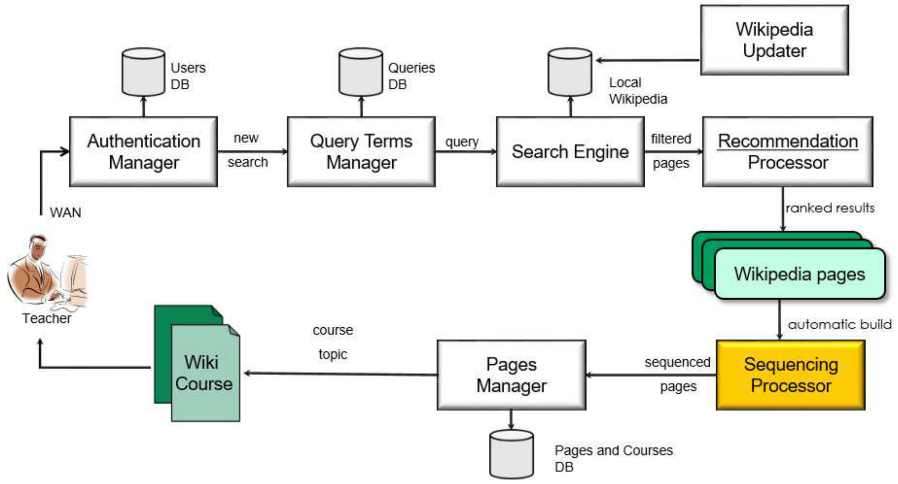
The second part of the Ph.D was dedicated to the development of the web platform for the creation of courses by teachers and their automatic sequencing using the classifier developed and validated in the first years of the Ph.D. Wiki Course Builder (WCB) allows teachers to find and automatically organize learning content exploiting Wikipedia. As it will be described in the next chapter, the teachers can select the language for the course and consequently use the right version of Wikipedia. The multi-language module manages the text analysis of the 13th most used languages in the world, but has been structured so that is easy to add support to other languages using external plug-in. Obviously the amount of materials found is dependent on the language used, so is recommended to use the English version that can count on 5,500,000 entities. The Wikipedia versions are locally stored as explained later and are updated twice a month via official dumps released by the WikiMedia Fundation. The final goal of the platform is to help teachers with a system that follows the users in the process of creating online courses, finding articles related to the topics and suggesting optimal training paths.

4.1 Wiki Course Builder Architecture

Wiki Course Builder [GLS15] consists of seven modules reported in Fig. 4.1

. The following sections describe the main modules and their functioning, the teaching model based on the teaching styles and the recommendation system.

Figure 4.1: WCB Architecture



4.1.1 Authentication Manager

The login manager manages the access to the platform. Enables users to register by email and takes care of initial user profiling that is addressed to a Grasha and Riechmann test for the initialization of teaching styles [A.G96]. The test consists of 40 questions to which the user must answer in a 1-5 scale. As a result a quintuple containing the values for each one of the figures described by the model Grasha-Riechman will be generated.

4.1.2 Query terms Manager

The module is dedicated to the management of the queries, with any contextual terms defined by the teachers; the QueryTermsController exposes to the GUI the methods necessary to obtain information, such as, for example, obtaining the suggested queries similar to the search keywords that the user is entering. Also here, the service layer is dedicated to the implementation of business logic, supported by the classes dedicated to dialogue with the persistence layer.

4.1.3 Search Engine

The module deals with the research of the initial concept for the creation of the course by exploiting the result of the query manager. The system presents the 100 concepts closest to the research according to 4 metrics that will also be used in the sequencing process of materials by the Sequencing Processor. The user can sort the materials according to each one of the 4 criteria and select the root concept of his course. In order to optimize the performance of this phase and the sequencing phase, we decided to have the entire Wikipedia in multiple languages locally and migrate to a no-sql database (MongoDB 2.7).

4.1.4 Wikipedia Updater

Wikipedia Updater deals with the management of the update of the various Wikipedia versions two times a month through the analysis of official dumps released by the WikiMedia Fundation. The dumps are provided in a very large xml format (Wikipedia eng = 65GB). The module is structured to run on multiple threads and uses the javascript tool dumpster-dive[1] to parse the Wikipiedia markdown language and generate an ordered structure. In addition, a pre-processing phase of the pages is applied in order to optimize performance at runtime; specifically, the stop-words are eliminated and algorithms of stemming and POS tagging are applied. In addition, the structures are prepared with all links and Wikipedia categories for the semantic analysis of the pages. Everything is saved in a collection of MongoDB on which a key index is generated with the page titles. The introduction of this no-sql database has brought great advantages in terms of execution speed in two processes:

(i) Research phase: with just one query it is now possible to discover all the concepts associated with the keywords selected by the teacher

(ii) Building phase: the time necessary to generate the Wikipedia graph in the construction phase of a topic has passed from the order of minutes to that of seconds. This result is very important because one of the criticisms made to the system was the excessive amount of time need for the creation of a course.

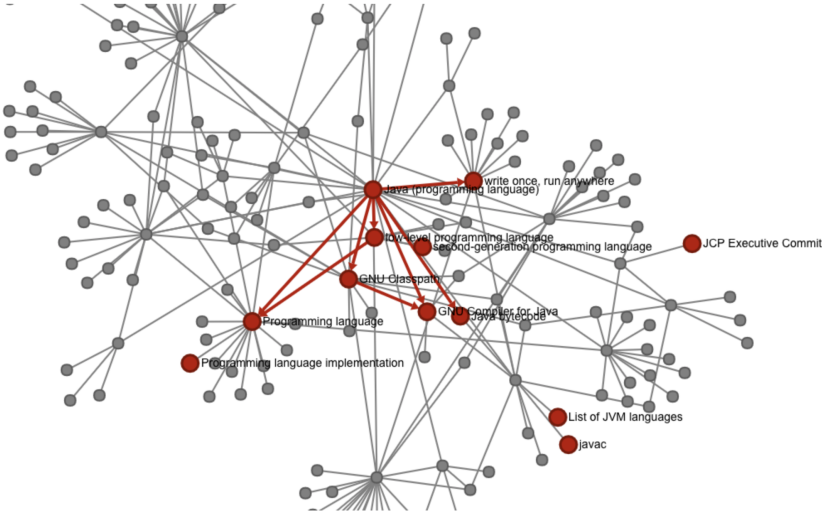
4.1.5 Recommendations Processor

The recommendations processor deals with the retrieving and the ranking of the Wikipedia pages. Starting from the page selected by the user during the search phase, the module extracts a set of Wikipedia pages associated with the main concept. The pages are

[1] https://github.com/spencermountain/dumpster-dive

selected by exploring and analyzing a portion of the Wikipedia graph. The Wikipedia Graph is the graph built from a Wikipedia page and taking all the linked pages; each node represents a page and the concept associated with it, the arcs are the links between the pages (Fig4.2). The pages are represented by bag-of-word after a stemming

Figure 4.2: portion of Wikipedia graph



process and are passed through four algorithms to generate four metrics. Three metrics are taken from classical literature and are TF-IDF, LSI, IG.

Term Frequency-Inverse Document Frequency (TF-IDF [Ram03]) is a mathematical formula for calculating the importance of a word inserted in a document in relation to other similar documents. This value is used as a weighting factor in information retrieval. The TF-IDF value increases proportionally to the number of times a word appears in the document, but is compensated by the frequency of the same word in the body of other similar documents, this helps to check if some terms are generally more common than others.

Latent Semantic Indexing (LSI) was used for internal research in archives and databases. LSI considers two documents semantically close if they have many keywords in common, while they consider them semantically distant if they have few keywords in common. LSI is based on the principle that words used in the same context tend to have similar meanings. A fundamental characteristic of LSI is the ability to extract the concept expressed in a text by creating associations with the terms that

are needed in documents that deal with similar contexts.

Information Gain (IG) measures the amount of information obtained for prediction of a set of documents *D* by knowing the presence or absence of a term *t* in a document *d*. IG is a measure of how common a feature is in a particular document d compared to how common it is in all *D*. In text classification, for example, given two terms in the corpus - term1 and term2, if term1 is reducing entropy of the class variable by a larger value than term2, then term1 is more useful than term2 for document classification in this example.

The fourth metric was created ad-hoc and it is based on the user model. An important point in the experimentation that will be presented in the next chapter was to demonstrate that the metric constructed in this way performs better than the metrics in the literature. The final result is a data structure containing all the most useful pages for the teacher ordered according to the four metrics.

4.1.6 Sequencing Processors

The Sequencing processor performs a double task:

- in the course creation phase it calculates logical paths highlighting them in red above the Wikipedia graph as shown in Fig.4.2. However, the disconnected nodes of the graph are suggested concepts but not connected with others.
- in the organization phase of the course deals with the communication with the binary classifier. Using the results of the predictions for all the pages of the course, the module presents to the teacher a conceptual map of the course with the prerequisite / successor relationships between the modules. This allows to have an innovative method of fruition of the courses that in all other platforms is essentially sequential.

4.1.7 Page Manager

The page manager takes care of saving the courses. Initially it saves the pages found by Wikipedia, but with the transition to the no-sql db the page manager now only keeps track of the last pages used in order to optimize the execution time of the back-end functions working like a cache memory.

4.1.8 Wiki Page Linker

This module was developed as a distinct unit; it is a web-service written in Java that can query the built model for the prerequisite / successor relations recognizer. The ser-

vice accepts incoming HTTP requests containing two wikipedia page ids. Here is the exemple of a request: // http://roma3ailab.it:8088/weblinker?id1=xxxxx&id2=yyyyyy

The processing takes place in the following steps:

- All the features described in 3.2 for the two pages are extracted.
- The features related to the page pair are calculated and the instance for the classifier is created.
- With the WEKA API, the model is interrogated to obtain predictions.
- The response is sent back to the server as JSON containing the value related to the found relationship.

The results of this processing are series of predictions on wiki-Pages pairs associated to the nodes of the course. This information will be combined to build the conceptual map of the course by analyzing the relationships found. In this way the teacher will be able to get the course sequenced automatically in order to optimize the learning process of the students. This process remains independent from the user model.

4.1.9 Classification of Teachers in Wiki Course Builder

Starting from the teaching styles described before, it is clear it is not possible to give a rigid categorization of a teacher in just one of the models; therefore each teacher is represented by a mix of these five categories representatives of their educational peculiarities. In order to represent each educator according to the particular mix of teaching styles that prefers, Grasha and Riechmann [Gra94] have written a questionnaire, consisting of 40 questions, to be submitted to each of them. For each question the teacher can answer with an integer value between 1 and 5, where 1 corresponds to fully disagree and 5 to fully agree. The questions are articulated according to statements, such as: - "My objectives and teaching methods address a variety of learning styles of the students", or -"Take the time to consult the students on how to improve their work in individual projects or group". At the end of the survey, the average of the answers to the questions of each group is processed and, for each of the 5 teaching styles, a decimal value between 0 and 5 is provided, indicating the level of the teacher in that particular category.

Fig 4.3 gives the scores assigned to each level (Low, Moderate, High) of each style.

Once the user is correctly profiled from the platform with the mix of teaching styles seen in the previous section, it is necessary to understand how to exploit this

Figure 4.3: Thresholds of teaching styles relevance

Stile	Low	Medium	High
Expert	< 2.8	<= 3.8	<= 5
Formal Authority	< 1.9	<= 3.1	<= 5
Personal Model	< 2.8	<= 3.4	<= 5
Facilitator	< 3.0	<= 4.0	<= 5
Delegator	< 1.8	<= 2.8	<= 5

information to personalize the search results. In the development of this section, we make the assumption that teachers with similar teaching styles might want to choose "similar articles" for their courses. Therefore, the platform keeps track of all the choices made by the teachers, proposing to them the search results as much as possible accordining to their teaching style. A social network is defined as a group of people linked by social relationships. The set of users of the WCB platform constitutes a social network, in which, currently, the link between users is determined by the teaching styles and the Wikipedia pages used in their courses. In the previous paragraphs the Grasha method was introduced for modeling users based on teaching styles; in order to exploit this modeling to submit personalized results to each teacher, it is necessary to define similarity metrics among the users of the system, which allow to group them in clusters with similar characteristics. Social Network Analysis [Zha10] is a methodology that allows to trace and measure relationships and flows between people, groups, organizations, web pages, producing a visual and mathematical analysis of relationships. In social networks, nodes usually represent people or groups, while arcs show relationships or flows between nodes. The Cluster Analysis introduces techniques that allow the identification and management of groups of users with similar characteristics. Therefore, a preliminary study of these techniques was carried out, in order to determine the metric to be used in the Wiki Course Builder to calculate the degree of similarity between users and, consequently, to order the articles extracted from Wikipedia favoring those chosen by similar teachers. Cluster Analysis [SMD+10] is the process of classifying objects, with the same or similar properties, in groups that are generated according to specific problems. These groups of objects are called clusters. Clustering is applied in different areas of research, such as biology (genetic engineering, neural networks), classification of information (web pages, text documents, search engines), climate (atmospheric and ocean patterns), psychol-

ogy and medicine (types of depression, spatial and temporal distribution of diseases). In Social Network Analysis (SNA) we can think of the cluster as a group of people with similar attributes (based on their interaction) or similar behaviors. The process in which group partitioning is sought for groups of objects, in which there are mutually similar objects, is called clustering. In an ideal situation, this process should meet two objectives: correctness and effectiveness. The criteria for correctness are as follows:

- Methods must remain stable as collections grow or, in other words, cluster distribution does not have to change drastically with the addition of new objects.
- Small errors in object descriptions must be reported as minor changes in cluster distribution.
- A method must not be dependent on the initial ordering of its objects.

4.1.10 Use Case

The Wiki Course Builder platform can be reached at the address

$http://roma3ailab.it:8080/login.html$

At the first access a simple registration by e-mail is required Fig 4.4. At the first

Figure 4.4: Registration module

Figure 4.5: WCB Homepage

access the user must answer the questionnaire on teaching style.

Each question is associated with a didactic role in the Grasha-Riechmann model and drives the system for the user model initialization. Wiki Course Builder works on the course-topic-material model. Each course consists of several topics and each topic is described by the Wikipedia pages necessary to learn it. From this moment, it is possible to start building the first topic. By clicking on the start button, the users can perform the first search. Is necessary to provide an initial concept for the course and some terms to disambiguate its meaning. From the dropdown menu the users can select the search language; this decision will influence the language of the entire course. The users can view the various pages obtained from the search and when they finds the most appropriate page by clicking on it they can go to the topic creation phase.

The topic building page is the heart of the system and contains all the features for creating the topic. Specifically it is divided into three sections :

(i) at the top users can specify if the current topic will be inserted in an existing course or if the system will have to create a new one; it also allows users to define the name for the topic that they are going to create. (see Fig. 4.7)

(ii) the second panel presents all the suggestions obtained from the recommendation processor. The user can see the found pages for each metric and add others through the Wikipedia graph by clicking on the "show graph" button. Clicking on a node in the graph highlights the links with the other nodes and it is possible

Figure 4.6: Search

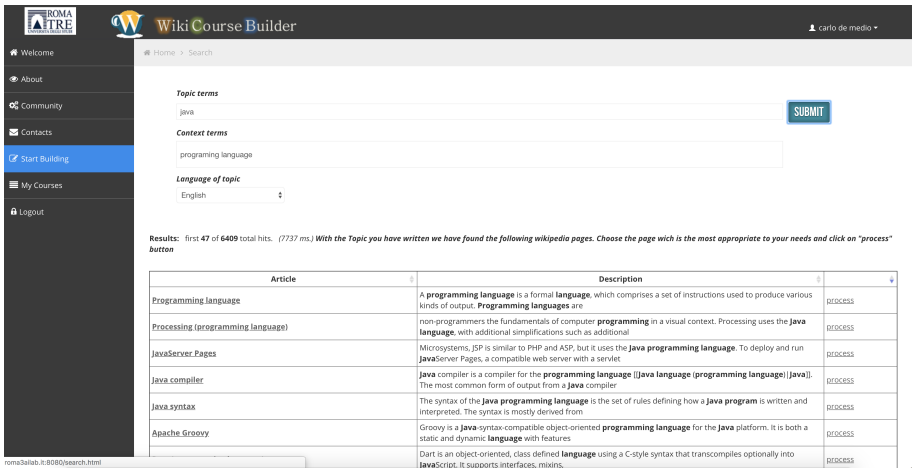
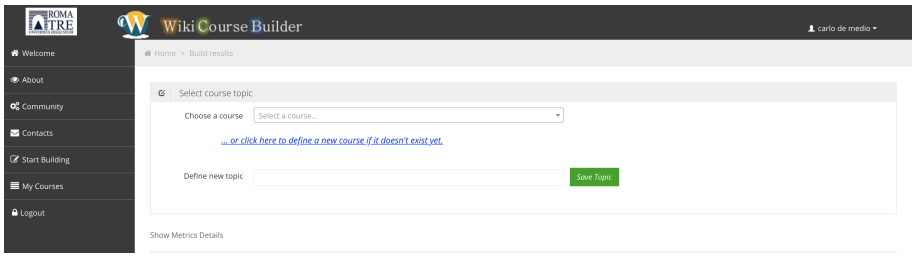


Figure 4.7: First section



to add the associated Wikipedia page by clicking on the green button add page to list Fig.4.8.

(iii) On the bottom of the page, users can rearrange the materials and in case they realizes to have done a mistake, delete them Fig.4.9.

When the topic is complete the user can click on the save topic button.

The user is redirected to the topic configuration page. The functions for adding pages are the same as those found on the construction page. Looking at the graph, the suggestions are no longer shown, but in red is highlighted the teacher's didactic path. Furthermore it is possible to export both the topic and the whole course in pdf

Figure 4.8: Second sections

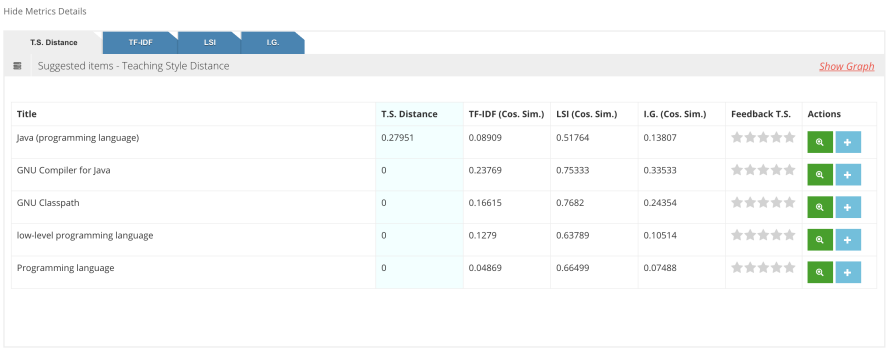
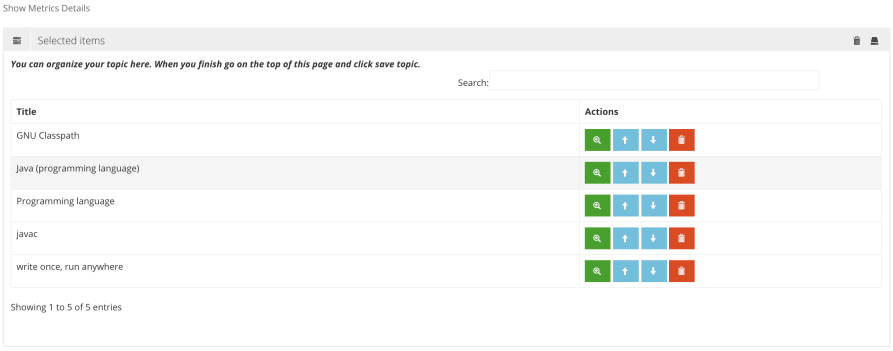


Figure 4.9: Third section

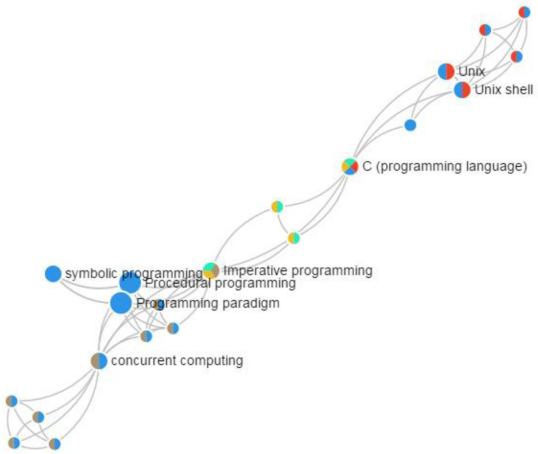


format. The result is a pdf file containing the teacher's data and a reconstruction of the selected pages with the possibility of selecting various options for the final layout, including the images. Finally under the graph the system proposes the conceptual map calculated through the interaction with the classifier ready to use.

From the main menu there are other important functions. The first function, accessible by clicking on my courses allows the users to see all the courses created by themeself and modify various topics. By clicking on the community, the users can compare themeself with the teaching community. An example of a community graph is shown in Fig.4.10 and shows all the courses in the system, colored according to the values of the teacher user models. The graph can be filtered by topics and allows the

users to interface themselves with the community favoring the sharing and reuse of materials.

Figure 4.10: An example of a section the Community graph



4.1.11 Experimentation

A large-scale experimentation was prepared to validate the quality of the user-modeled recommendation system. In order to overcome the problems associated with the cold start, teachers from Roma Tre were recruited to use the system and create courses in various research fields. Afterwards, a call was organized for the teachers on the MIUR SOFIA [2]http://sofia.istruzione.it/) platform so as to be able to reach professors from all over Italy and at different levels. The Sofia platform was developed to allow accredited organizations (junior high schools, high schools, universities ...) create online courses; teachers can propose or follow training initiatives; the platform also issues certificates for training credits.

The experimentation plan in detail is presented below; the main questions are:

[2](

(i) Is the metric based on the teacher model for the selection of pages from Wikipedia better than the metrics based only on the contents of the pages?

- Grasha model (already verified)
- Selection of the metric
- TF-IDF, IG, TS, LS

(ii) Does the management of the contents map directly made by the user by viewing the graph give added value to the preparation of a course?

- Happy sheet.
- Implicit feedback.

(iii) Are the courses prepared with the system system better compared to courses prepared without the system?

- Experiments on students in classrooms with control groups
- Anonymous questionnaires
- Comparison of evaluations with users coming from traditional courses

All teachers used the system for 16 hours creating courses on any subject. All human-machine interactions were logged and analyzed. In particular, teachers were asked to perform 2 simple tasks:

(i) in the creation phase of a course, we asked to evaluate the materials found by the system on the 0-5 scale. As it is shown in 4.8, users could not know from which system metric the suggestion came from guide the answers.

(ii) to take a picture of the conceptual map of the materials included in the courses, written by hand or as they wish and send it via e-mail.

A questionnaire was prepared to validate the quality of the courses provided by the platform; in addition to the implicit feedback left within the system, each teacher had to answer 10 questions to evaluate the courses created. In future developments it would be good to consider the idea of experimenting with the students; dividing the sample of students into two groups it would possible to see the results of the final tests of groups educated with the classical courses and courses built with WCB. From the results it will be possible to evaluate the effectiveness of the teaching paths generated automatically.

4.1.12 Evaluation

The experimentation on the Wiki Course Builder platform aimed to validate 2 things:

(i) Demonstrate that the ability of the system to rediscover materials from Wikipedia and to suggest didactic paths based on the developed metric is better than that obtained by following the classical metrics of the literature.

(ii) The capacity to provide teachers with a tool that would improve their experience of creating online courses, simplifying their processes and speeding up the entire task.

For this experimentation we show the results of the evaluations made by analyzing the implicit feedback given by the teachers to the recommendations obtained. It was necessary to consider the well-known problem of cold start affecting all user-modeled systems. Initially, the materials have no indications on the proposed metric so initially the best recommendations come from text-based metrics.

At the end of the experimentation, two anonymous questionnaires of ten questions were administered to each teacher; the first one is structured to quantify the usability of the platform and the second is classic Happy sheet to capture user satisfaction. The questionnaires are hosted on the GOOGLE cloud.

In this first part of the evaluation we evaluate not all the system but its capability to retrieve Wikipedia pages, compatible as much as possible with the teacher model, i.e., with the teaching styles of the instructional designer who is building a new topic to use into one or many old or new courses. We show an evaluation of the didactic TS-metric Vs. the other three content-based metrics, discussing the experimental results. The research question to discuss is: are the Wikipedia pages, retrieved by using the TS metric, more appropriate for the teacher who launched the query with respect to those pages retrieved by using the other three content-based metrics? Content-based metrics are important for overcoming the so-called *cold start problem* while, the TS metric will be the only one able to make a contribution to the didactic of each teacher after a time depending on the use of the system: the more the system will be used, the more the TS metric will overcome the other metrics. In fact, every time a given material is retrieved and used by the teacher, that Wikipedia page is tagged taking into account the models of the teacher who used it.

4.1.13 Data Gathering

The sample is composed of 26 teachers, whose models are represented in Tab. 4.1, coming from all research area who retrieved on average 4 topics each for a total of

Table 4.1: The average teacher model.

Feature	Mean	SD
Expert	3.52	0.60
Formal Authority	3.84	0.42
Personal Model	3.6	0.71
Facilitator	3.7	0.32
Delegator	3.6	0.37

Table 4.2: The Statistics of the sample.

	TS	TF-IDF	LSI	IG
μ_x	3.2	3.0	2.5	2.8
σ_x	1.48	1.55	1.6	1.56

104 topics. Each teacher built a new course on his study field. During each work session, each teacher assessed all the retrieved materials proposed by the system for each metric by means of a rating from 1 to 5. Each teacher rated at least 20 items for topics, 4 metrics for each retrieved topic, formed by almost 5 pages, building a database of 3000 ratings. After each assessing session, each teacher was required to fill-in two questionnaires: the first for the usability and the second as happy-sheet, in order to gather indications to improve the system from different points of views. The statistical data are shown in Tab. 4.2, where for each metric the arithmetic mean of the sample, μ_x, and the standard deviation of the sample, σ_x are reported. As one can see, the standard deviations are almost equal among them while there are some substantial differences among the arithmetic means.

4.1.14 The Statistical Test

In order to check our research question, we use the *t-test* as the statistical test to verify the differences between the TS metric Vs. each other one. So we run three statistical t-test, one for each couple. The sample is formed by 50 topic assessed by the teacher for each metric. So we have:

- Hypothesis H_0: there is no difference between the distribution of ratings given to the TS metric Vs. the one given to the other metrics;
- Hypothesis H_1: the distributions to which belong the two sets of ratings are different and the TS metric is greater than the others;

As p-value we choose $\alpha = 0.05$. The statistical results are shown in Tab. 4.3.

Table 4.3: The t-test.

	p-value	H0	H1
TS vs. LSI	0.0276	Rejected	Accepted
TS vs. IG	0.045	Rejected	Accepted
TS vs.TF-IDF	0.53	Accepted	Rejected

4.1.15 Discussion

From Tab. 4.3 we see that the TS metric belongs to different statistical distributions with respect to the LSI and IG metric. For these two metrics, the H_0 hypothesis can be rejected and the TS metric being the p-value less than α. Furthermore, the TS mean is greater than the other two means. For the TS Vs. TF-IDF case, the H_1 hypothesis is rejected, being the p-value greater than $alpha$. In our opinion, the reasons of these results are, in the first case, due to the nature of IG and LSI metrics. Both these metrics work on words correlations and, in order to index documents, need a huge collection while WCB retrieved 5 pages each time. The TF-IDF metric works on terms occurrences in documents and, until the system will not have a huge database of retrieved and used pages it will be like the TS metric. Unfortunately this is a problem connected to the cold-start and we used this metric to overcome this problem.

4.1.16 The User Feedback

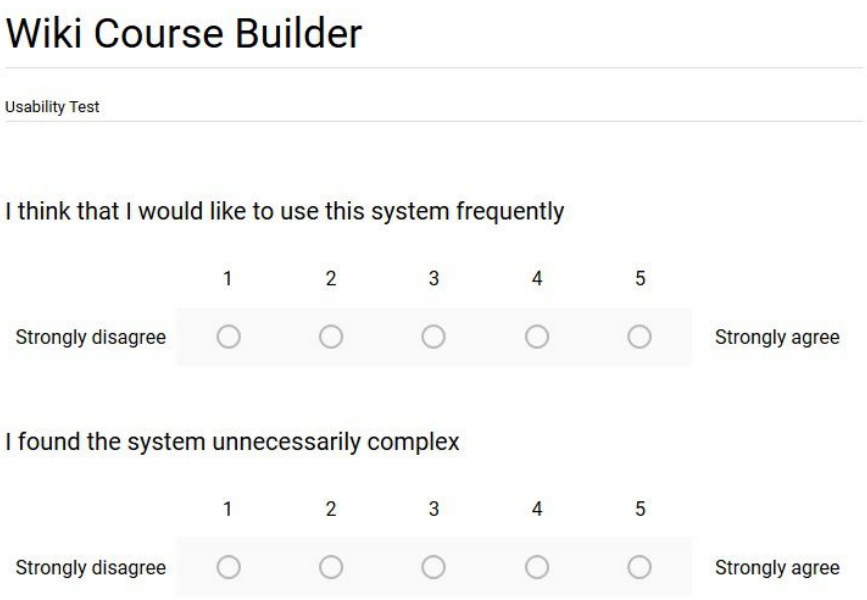
In order to have an overall evaluation of the system, each teacher, after having finished her session filled-in two questionnaires. The first questionnaire, composed of 10 questions, concerned the usability of the system while the second, composed of 10 questions too, was a happy-sheet to have measure of the satisfaction degree in the use of the system itself.

4.1.17 The Usability Questionnaire

This questionnaire is the *SUS usability questionnaire* (Fig. 4.11). After having built a new course, in the home page two new icons appear. The user, clicking the *Usability* icon is redirected to a 10-questions questionnaire built in the google-module environment. The questions are 5-likert scale ones, taken from the SUS. We obtained

as $\mu = 65$ which is a good rating. [3]. For a deep insight the reader can read "Measuring Usability ith the System Usability Scale [4].

Figure 4.11: A part of the Usability SUS questionnaire



The Happy Sheet

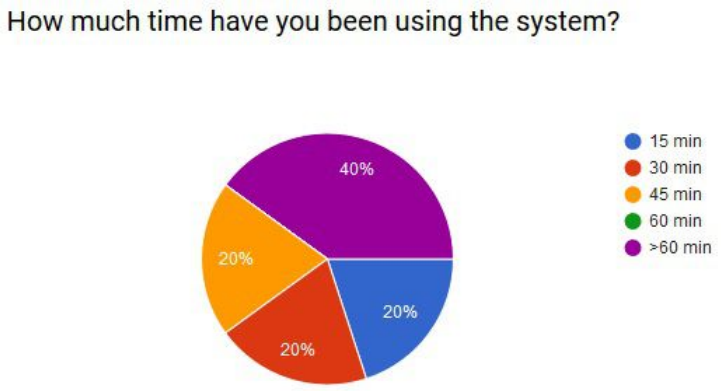
This questionnaire is composed by 10 questions with the aim of measuring the user's satisfaction degree in the use of the system. This questionnaire is available on-line.

4.1.18 Conclusions

WCB is a system presenting several features, from the retrieval of Wikipedia pages directly from Wikipedia, to the creation of a new course as a set of topics. An important feature of the WCB system is its capability to model teachers by means of their teaching styles, basing on the Grasha Teaching Styles model. Each teacher is first

[3] https://www.usability.gov/how-to-and-tools/methods/system-usability-scale.html
[4] https://measuringu.com/sus/

Figure 4.12: Results



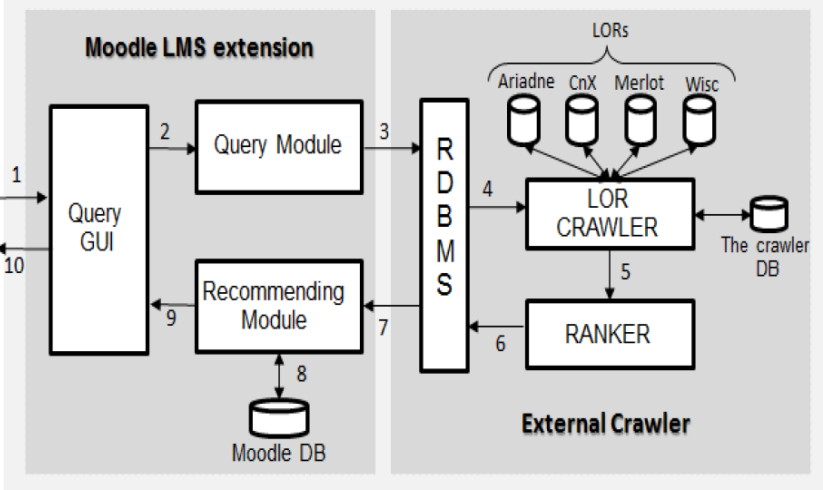
modeled by five dimensions, expressed as a set of real values. Another characteristic of the system is the modeling of those Wikipedia pages already used by other teachers belonging to the community using the system: each page is marked by a 5-dimension array representing the teachers who used it in the past. We presented an evaluation of the first step of the course creation process, i.e., of the retrieval phase. We addressed the research question to check if the retrieval metric based on the teaching styles retrieved better Wikipedia pages, i.e., Wikipedia pages more adapted to the teacher who launched the query. We conducted a first experimentation with positive results. As a future work we plan to conduct a larger evaluation of the overall system.

Chapter 5

Moodle_REC

As mentioned previously, the retrieving of teaching materials is one of the tasks that consume more time in the process of course creation. The strongest hypothesis that is made in the presentation of the Wiki Course Builder platform is that all the topics are composed by only Wikipedia pages. The accuracy of Wikipedia was defined as comparable to that of the British Encyclopaedia in a 2005 study [Gil05], but different teachers may wish to use materials constructed differently from the classic html pages. Also using only Wikipedia we are not considering the reuse of all materials produced by teachers on the web and labeled in online repositories. Furthermore, it is not possible to do a material analysis as described in Bloom's taxonomy [For10], since all the pages have the same structure and it is not possible to identify significantly different learning styles. These motivations have led to the development of an external tool for the efficient retrieval of Learning Objects from the most famous repositories, MoodleRec [LLM+15] and in [DMGL+17b]. The extension was integrated into Moodle LMS to test the goodness of the recommendations; it was decided not to integrate the extension into Wiki Course Builder in order to avoid, in the results, the noise caused by the lack of knowledge of the platform. Moodle is a platform used in schools of all levels and it was assumed that the teachers were more practical. The extension can easily be integrated into all Learning Management System since it has been developed as an external web-service; through simple http calls it allows to obtain a sequencing of the materials found on the web.

Figure 5.1: Moodle Rec architecture.



5.1 Moodle Rec Architecture

In this Section we show the architecture of the system. After a brief description of the overall architecture, depicted in Fig. 5.1, we focus on the *Recommending Module*, which is the main component of our work. The Recommending System (RS) is composed of two main modules. The first one is embedded into the Moodle LMS, extending some functionalities, while the second one, acting as the crawler system, is an external web-based system, as shown in Fig.5.1. The two modules interact by a *RESTful* web-service communication architecture[1]. In the following subsections we describe the functional features of each module.

5.2 The Moodle LMS Extension

One of the focal points of our approach is that the user can use the system directly into the Moodle LMS environment.

The Query GUI. We have extended the functionality *add Url* within a section. Launching *add url* the user accesses the Query GUI that allows the teacher to set both

[1] https://www.tutorialspoint.com/restful/

the search parameters or keywords associated with the LOs to be retrieved from LORS, and the LORS on which perform the search, among those permitted by the system: *Ariadne, Merlot, CnX* and *Wisc*. The GUI also shows the results of the Recommending Module and implements the User Interface for exploring the recommendations.

The Query Module. It receives in input the query transmitted by the Query GUI and after a query expansion process communicates with the external RDBMS via web-service. The external system will be explained in the next sub-section.

The Recommending Module. After receiving the external service output, it shows the environment by which one can manage the retrieved LOs. By this module, one can select those LOs assessed as relevant for the course at hand and subsequently select other LOs basing on the recommending algorithms and metrics explained in the next section. This module communicates with the Moodle_DB, and it shows a set of features related to the materials used in the system. It also performs the recommending strategy starting from the teacher model, recommending to the teacher, among those LOs already used by other teachers, those LOs assessed as to be inserted in the course together with the selected one.

The Moodle_DB. This database contains, in addition to the Moodle records, all the information needed by the RS. It also logs all the steps for the future experimentations.

5.3 The External Crawler

As shown in Fig. 5.1, the External Crawler is a module with the aim of communicating with the selected LORs. It is composed by three modules that are described below.

The Crawler. It extracts the LOs from the repository on line and stores the information; this module stores the retrieved LOs in a MYSQL database, acting as a sort of local LOR. In fact all the retrieved LOs from the crawler are stored locally in order to speed up the retrieval time for similar queries. The crawler updates the LOs data every night.

The Ranker. It combines the information retrieved by the Crawler from the repository with the *TF-IDF* measure in order to give a normalized score for every material; this measure is used in the recommendation module to give a possible order for the LOs.

The RDBMS. This module plays a crucial role because it retrieves from the *Crawler DB* (see Fig. 5.1) those LOs assessed as relevant for the user query; this module implements the communication protocol between the Moodle Extension and the External Service DB.

5.4 The Recommending Strategy

In the system a course is representable as a sequence of LOs:

$$C = \{lo_1^C, lo_2^C, \ldots, lo_{n_C}^C\}$$

In the above C, given h, k, both $\in [1 \ldots n]$, with $h < k$ then we say that lo_h^C *precedes* lo_k^C with *distance* $dist(lo_h^C, lo_k^C) = k - h$. The interface allows to select a value d for the distance. The default is 1, that indicates an immediate successor or predecessor of the given LO (l). Let's assume that the teachers $T = \{t_1, t_2, \cdots, t_{n_T}\}$ is the set of all teachers working in the system. And let $C(t) = \{C_1^t, \cdots, C_{n_t}^t\}$ be the set of courses managed by t in the system. So the $i-th$ course of the teacher t would be $C_i^t = \{lo_1^{C_i^t}, \ldots, lo_{n_{C_{i,t}}}^{C_i^t}\}$.

Information about a retrieved LO: basic relevance and usage count. After a query the teacher is shown a list of LOs. Let's assume that the teacher highlighted a LO l from such a list: then, the first information shown by the interface is the number of occurrences of l in the courses managed in the system. This allows the teacher to evaluate how much l has been already used by her colleagues.

Information about a retrieved LO: similarity of didactic contexts. On the other hand, a relevant information to show, is the context in which l has been used in such courses i.e.:, what are the other LOs that are presented together (before or after) l in other courses.

To such aim we define the following operations, where d is a distance coefficient, chosen by the teacher to state how deep the didactic context analysis should be:

pred$(l, d, C) = \{l' \in C : dist(l, l') = d\}$.
This is the set composed by the predecessor, at distance d of l in C, or the empty set if there is no such predecessor.

succ$(l, d, C) = \{l' \in C : dist(l', l) = d\}$.
As above, this just regards the successor, at distance d of l in C.

PContext$(l, d, C) = \{pred(l, d, C) \cup pred(l, d-1, C) \cup \cdots \cup pred(l, 1, C)\}$.
This is implemented as a sequence, in the union order, from the farthest predecessor to the immediate predecessor of l in C.

SContext$(l, d, C) = \{succ(l, 1, C) \cup succ(l, 2, C) \cup \cdots \cup succ(l, d, C)\}$.
This is implemented as a sequence, in the union order, from the closest to the farthest successor of l in C.

$D(l, d, C) =< PContext(l, d, C), SContext(l, d, C) >$.
It is the *didactic context* of l's occurrence in C, bounded by the distance d.

We use the above operations to define the didactic context of a LO in a course, as the representation of what is around that occurrence: what other LOs precede it, and what follow it.

Then, given two courses, C, C' in which both l occurs, we may evaluate how similarly l is used in them, by comparing $D(l, d, C)$ and $D(l, d, C')$, and deepen such comparison at different distances d.

Eventually, also similarity of two teachers can be computed, basing on the comparison of the didactic contexts of the LOs they use in their courses.

We are trying to establish a framework for such comparisons, so we start by a naive definition of didactic context similarity: in particular, given courses C and C', a stated distance d and a LO $l \in C \cap C'$, the similarity of two didactic contexts, $CtxSim()$ is computed as follows. Let be
$PContext(l, d, C) = \{l^C_{i_d}, \cdots, l^C_{i_1}\}$,
$PContext(l, d, C') = \{l^{C'}_{j_d}, \cdots, l^{C'}_{j_1}\}$,
$SContext(l, d, C) = \{l^C_{h_1}, \cdots, l^C_{h_d}\}$,
$SContext(l, d, C') = \{l^{C'}_{k_1}, \cdots, l^{C'}_{k_d}\}$,
and be $equal(l, l')$ a function returning 1 if $l = l'$ or 0 otherwise, then

$$CtxSim(D(l, d, C), D(l, d, C') =$$

$$= 1 + \sum_{r=1}^{d} 2 \cdot equal(l_{i_r}, l_{j_r}) + \sum_{r=1}^{d} 2 \cdot equal(l_{h_r}, l_{k_r})$$

Briefly, $CtxSim$ is an integer number that measures how many LOs are common in the neighbors of l in C and C'.

For instance, if $PContext(l, d, C) = \{a, b, c\}$, and $PContext(l, d, C') = \{c, b, e\}$, then the related addendum in $CtxSim()$ would be $\sum_{r=1}^{d} 2 \cdot equal(l_{i_r}, l_{j_r}) = 2$ $(0 + 2 + 0)$.

Basing on $CtxSim()$, we can define an operational semantics of the similarity between two courses, C, C' defined as above, within distance d:

$$CrsSim(C, C', d) = \sum_{r=1}^{n_C} CtxSim(D(lo_r^C, d, C), D(lo_r^C, d, C'))$$

where n_C is the number of LOs in the course C. Briefly, for any lo in C, that is also present in C', we compute the distance and sum all of them.

The information shown to the teacher, t, in this section, given a retrieved LO l and a distance d is then

- the list of LOs that precede l, within distance d, in other courses in the system;
- the list of LOs that follow l, within distance d, in other courses in the system;
- each of the above LOs links to the courses where it appears (as predecessor or successor of l; so the teacher can inspect the didactic contexts of such occurrences;

Moreover the above analysis data can be focused limiting to teachers that are similar to t.

The teacher model. In order to rate the LOs occurrences also basing on the similarity between teachers, the system manages a *Teacher Model*, which, for a given teacher t, is the collection, $TM(t) = C(t)$, of all the courses managed by t in the system (that is by a collection of sets of LOs, where each LO appears in a specific didactic context).

Similarity of two teachers, t, t' is computed basing on the similarity of the didactic contexts of the LOs they both use in their courses. So, given the courses of teachers t and t', as defined above, $C(t)$ and $C(t')$

$$TchSim(t, t') = \sum_{C \in C(t)} \sum_{C' \in C(t')} CrsSim(C, C', d)$$

$TchSim(t, t')$ is normalized over the maximum similarity possible for the teachers.

The similarity is calculated at run time. When a teacher asks for the recommendation for a material the system searchs for instances of the material in all courses. Analyzing the context (selected by user) of all courses the recommendations with hight similarity level are reported.

5.5 Use Case

In the following example we will show only the more relevant aspects of the RS. Let us suppose that Emily Faraday wants to arrange a Moodle course about the introduction to Fractions and that in the System there are two other teachers, Alice Brown and Carl Dennison that already have two courses respectively on Pre calculus and History of Mathematics. Fig5.5 shows the two courses present in the System. Emily wants to add LOs: this actions is executed by adding the resource URL in Moodle. In Fig5.4 we can see that the usual Moodle page has been enriched with the section that allows the search in the repositories. She adds some introductory LOs: Introduction to Rational Numbers for Middle School Students and Integers, Rational/ Irrational Numbers. In the second section she wants to introduce Operations and Comparisons with rational numbers therefore performs the query operations with fractions comparison. In Fig. 5.2 the response of the system is shown . We observe that in first positions there are LOs with occurrence number ≥ 1 that means those LOs have been already used in the system. By selecting recommendations Emily can see in which context other teachers have used those LOs. She decides to add the second LO to her course. Then she performs another search with keywords least common multiple - see Fig. 5.3. and she finds the LO Arithmetic Review: The Least Common Multiple that occurs twice in the system. The details show that Alice Brown and Carl Dennison have used that LO, but the similarity with former teacher is 1, since they have another LO in common, while the similarity with the latter is still 0 (see Fig. 5.3).

.

5.6 Experimentation

For the experimentation an instance of Moodle was created and the extension was configured. The platform was left empty to check how the cold start problem was overcome by the proposed system. The teachers were recruited through the MIUR's SOFIA platform. Each teacher was asked to join the platform and create one or more courses on random topics. All interactions with the system were logged and in particular five data were saved:

- number of users;
- number of courses;
- number of recommendations seen by the teachers;
- number of recommendations followed;

- number of materials viewed.

Furthermore, as for the Wiki Course Builder platform, two questionnaires were prepared to evaluate user satisfaction and system usability.

For the purposes of experimentation, only the statistics calculated from materials of the external resource type were considered.

To verify the quality of the recommendations based on the teaching model, we recruited 25 teachers who had experience with the Moodle LMS. In order to validate the approach we have decided to integrate the extension in Moodle because in the usability tests the user's inexperience could introduce noise in the data. The extension has been designed to be compatible with all existing LMS. At the end of the experimentation, the teachers included in the platform a total of about 350 materials divided into 30 courses. 100 materials were selected starting from the recommendations, so about 33% of the materials. Fig.5.6 highlights how the growth of the pool of teachers influences the recommendations. Data shown in Fig. 5.6 are very encouraging because only few of the courses developed have arguments in common and therefore it can be assumed that in the future it will be possible to recommend an even greater number of elements.

.

Figure 5.2: The response of the system. On the bottom side the window with the information about the use of the second LO.

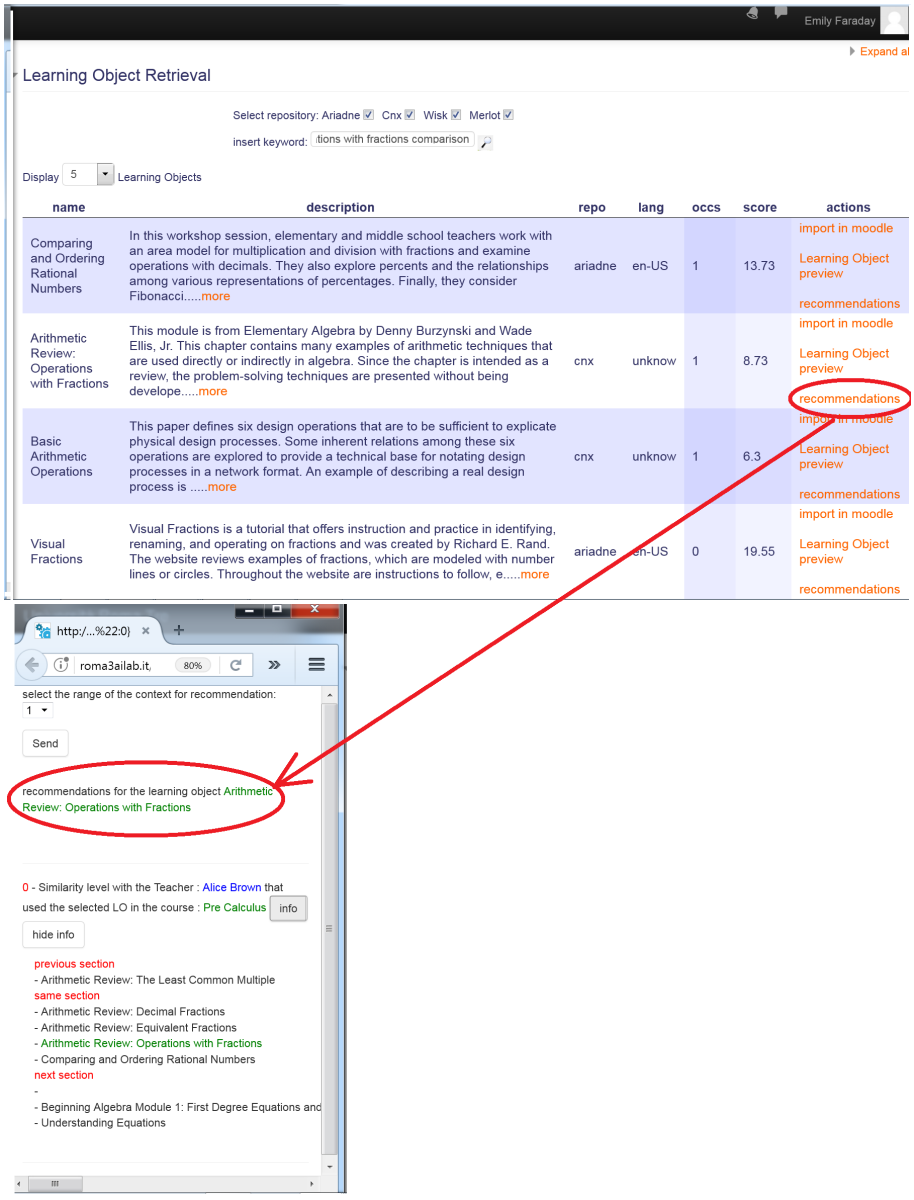


Figure 5.3: The response of the system. On the bottom side the window with the information about the use of the second LO.

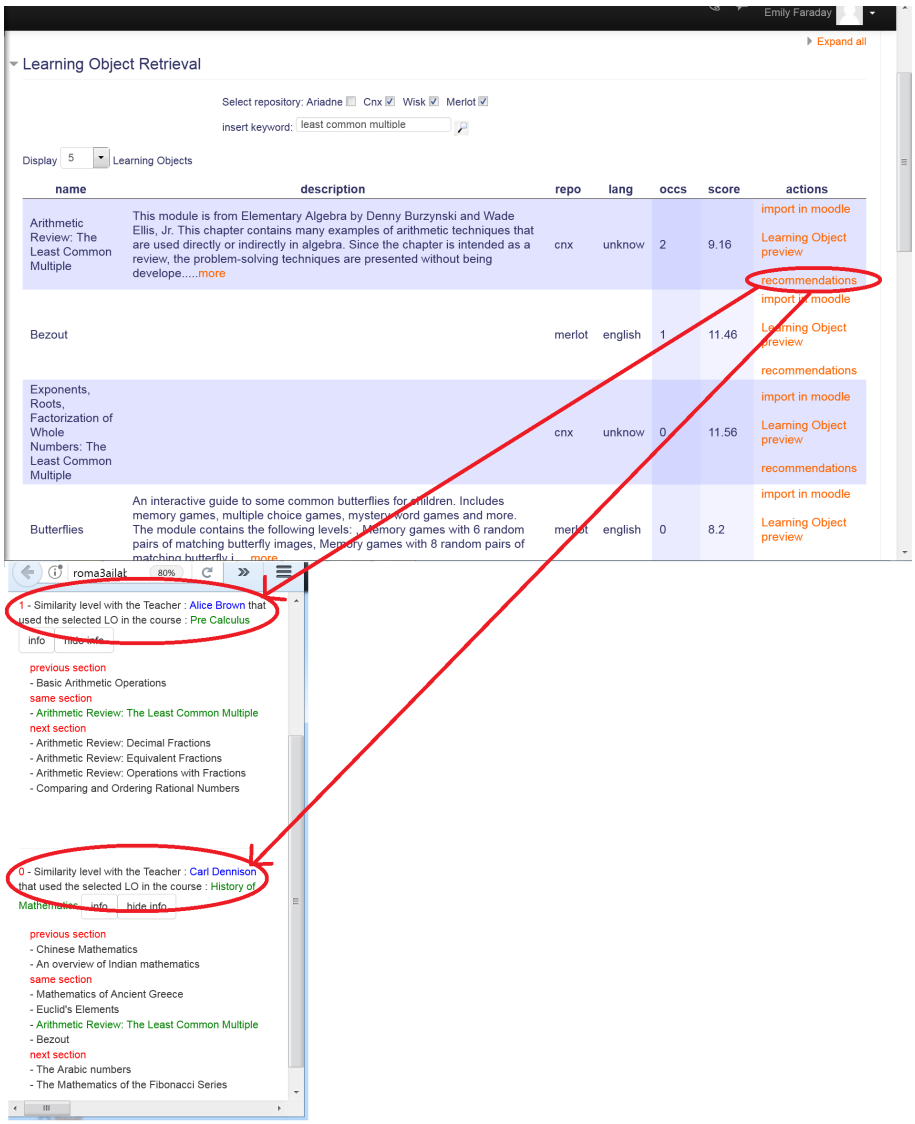


Figure 5.4: The interface for searching into repositories. LO.

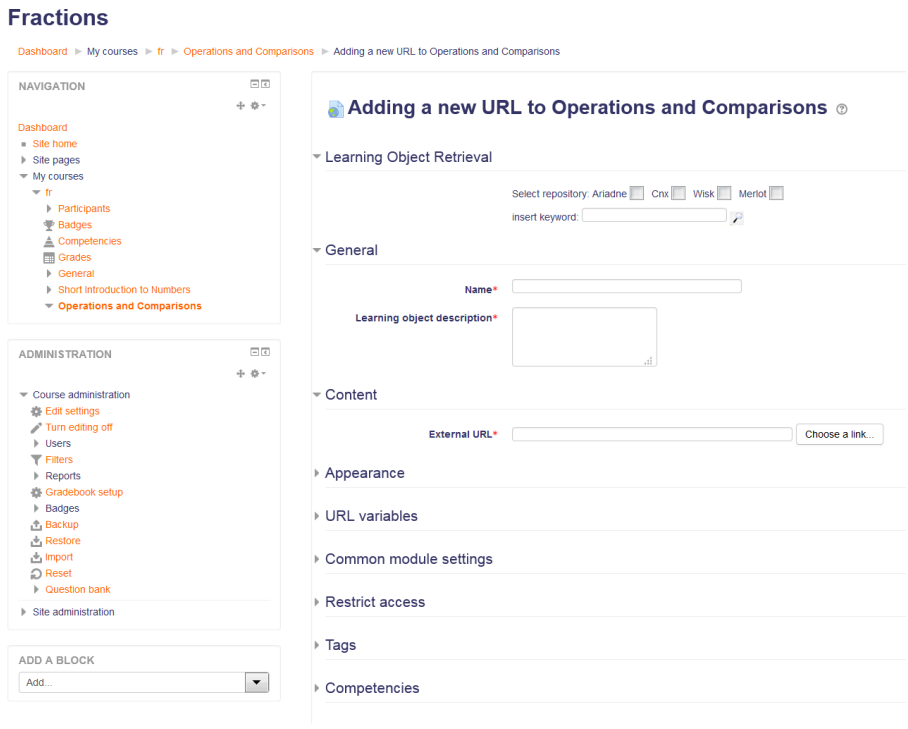


Figure 5.5: Course on History of Mathematics and Pre-calculus already present in the System. LO.

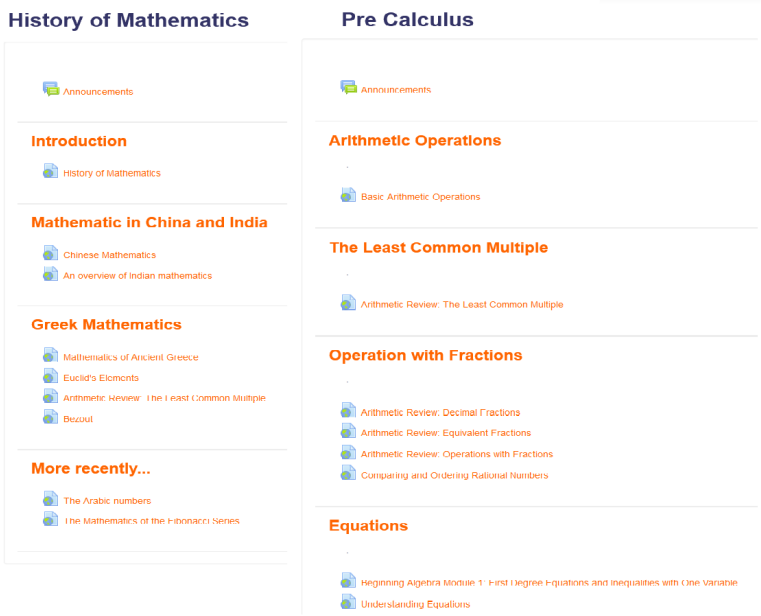
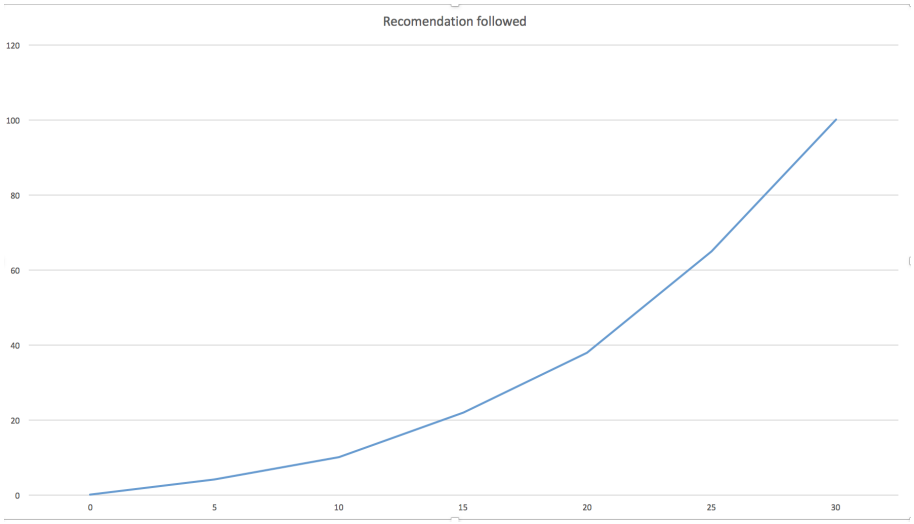


Figure 5.6: Results on Moodle Reco: on the x-axis the number of courses in the system and on the y-axis the reccomandations followed by the teachers.



Conclusions

Conclusions

In the three years of research the problems proposed in the introduction of this thesis were explored in depth. In conclusion, a platform has been proposed including all the functions to automatically create online courses, taking advantage of recommendations based on the described teacher model. The system allows users to find Wikipedia pages from internet and organize them in topics and courses. It also provides a graphical interface to simplify the material selection process. The community graph allows teachers to compare, and allows easy reuse of the pages within the courses. Moreover, through the prerequisite relationship recognizer, it presents at the teachers a pre-compiled course map and allows them to export the whole course in pdf format automatically.

The experimental results highlight the goodness of the approach and show how the platform can revolutionize the idea of online MOOC showing how it is possible to recommend didactic materials overcoming the classic use of metadata answering to RQ1. The results for the prerequisite/successor relationship are the the highest in literature, exceeding the previous approach by more than 30%. The results of the experiment on the binary classifier show that it is possible to identify the prerequisite/successor relationship between generic materials present on the web and validate the approach for the resolution of RQ2 using an innovative approach. It also enables the generation of concept maps automatically; predictions can be applied to course main concepts in order to obtain automatic study plans for students. Furthermore, the Moodle Rec extension has been presented for the Learning Object retrieval from the repository. The system has a crawling module that constantly updates the database of materials from the most used repositories. The module was written to be easily extensible to any new repository. The results of the experimentation highlight the goodness of the approach and respond to the RQ3 demonstrating that the teacher model must take account of the other teacher experiences.

Future works

For the future, it is possible to act on several points to improve the system's performance and extend its functionality.

Surely, a more in-depth semantic analysis can help to optimize all content analysis processes. An interesting study is [Poc17] in which the researcher tries to make an analysis of the critical thinking of the users. Starting from a set of indicators, they try to evaluate the qualities of the material. The current system is based on the exploitation of Wikipedia. Extending the approach to other online knowledge bases

would improve forecast accuracy. The difficulty in this case will be to be able to make the features homogeneous between the various bases of knowledge such as DBpedia [DBp].

Another interesting idea is to exploit other MOOCs repositories like Coursera to expand the training dataset starting from the intuition of [DMGL$^+$17a] [MGL$^+$16] where the authors analyze the sequencing of the materials to infer the relationships between them.

In order to simplify the creation process of the concept maps proposed to teachers, a graphic system can be useful. The tool created by the University of Genoa presented in [AK16] and [KM04] contains all the necessary functions. Users can add nodes and clicking on the nodes create an arc between them. A collaboration between the projects could lead to an integration tool for Wiki Course Builder. The integration was organized as an external module and presented in [DMGL$^+$18]. Further experimentation will be needed to evaluate the effectiveness of the tool within the system.

In the future, it will be useful to analyze the problem of the courses created by analyzing the results of the students; in literature there are different approaches but the most interesting are the following two.

(i) [WPP$^+$00] [ZLW$^+$17] analyze the results of the students' quizzes to define a set of features that characterize the quality of the given answers. These features evaluate the material produced by the student automatically. The measures are based on the calculation of n-grams of words and characters.

(ii) In [MLR$^+$18] the authors try to achieve the same goal by defining a new type of quiz that can evaluate the given answer through a set of rules.

Combining the intuitions of these two methods it is possible to develop an intelligent agent able to automatically evaluate any type of response to quizzes (multiple choice, short answer, ...).

The final system should propose connections to the teacher driving him in the concept map creation process. To define a more accurate user model, the similarity metric between the concept maps defined in [LLM$^+$16] can be integrated. Based on joint nodes and flows on the graphs representing the concept maps the authors can define the similarity between two maps. This can be reported as the similarity between the teaching styles of two teachers.

Also, it would be desirable to further develop the Community section, allowing users to interact with each other and share content; this aspect would favor the evolution of the Community of Practice and would increase the overall effectiveness of the system.

Appendices

Appendix A: List of Publications

Publications

This is the list of publications which have been generated from the work presented here. Other papers from this work are going to be published.

Journal Publications

- C. De Medio, F. Gasparetti, C. Limongelli, F. Sciarrone and M. Temperini Prerequisites between Learning Objects: Automatic Extraction based on a Machine Learning Approach Telematics and Informatics An Interdisciplinary Journal on the Social Impacts of New Technologies ELSEVIER. 2018.

Conference Publications

- C. De Medio, F Gasparetti, C Limongelli, F Sciarrone, M Temperini A Machine Learning Approach to Identify Dependencies Among Learning Objects 8-th International Conference on Computer Supporter,Education, april 2016.
- C. De Medio, F Gasparetti, C Limongelli, F Sciarrone, M Temperini Automatic extraction of prerequisites among learning objects using wikipedia based content analysis analysis,International Conference on Intelligent,Tutoring Systems, june 2016.
- C. De Medio, F Gasparetti, C Limongelli, F Sciarrone, M Temperini Mining Prerequisite Relationships Among Learning Objects , International Conference on Human-Computer Interaction, july 2016.
- C. De Medio, F. Gasparetti, C. Limongelli, M. Lombardi, A. Marani, F. Sciarrone and M. Temperini Discovering Prerequisite Relationships among Learning

Objects: a Coursera-Driven International Conference on Web-Based Learning, ICWL 2016, October 2016.

- C. De Medio, F. Gasparetti, C. Limongelli, M. Lombardi, A. Marani, F. Sciarrone and M. Temperini Analysis of Coursera Resources for Characterization of Educational Material on MOOCs The 2nd International Workshop on User Modeling for Web-based Learning (IWUM), Oct 2016.
- C. De Medio An Intelligent Agent with Ontological Knowledge: Classification of Educational Materials to Support the Creation of Online Courses ICWL 2016 Doctoral Consortium, Oct 2016.
- C. De Medio , F. Gasparetti, C. Limongelli, F. Sciarrone Automatic Extraction and Sequencing of Wikipedia Pages for Smart Course Building IV 2017.
- C. De Medio, F. Gasparetti, C. Limongelli, F. Sciarrone and M.Temperini Course Driven Teacher Modeling for Learning Objects Recommendation in the Moodle LMS In UMAP 2017.
- C. De Medio, F. Gasparetti, C. Limongelli, F. Sciarrone Modeling Teachers and Learning Materials: a Comparison Among Similarity Metrics In IV2018.
- C. De Medio, F. Gasparetti, C. Limongelli, F. Sciarrone, I. Torre, G. Adorni ,F. Koceva Wiki Course Builder, a System for Managing and Sharing Didactic Material and Concept Maps In EDEN RW10 2018.

Bibliography

[A.G96] A.Grasha. *Teaching with Style: A Practical Guide to Enhancing Learning by Understanding Teaching and Learning Styles*. Alliance Publishers, 1996. 10, 36

[AK16] Giovanni Adorni and Frosina Koceva. Educational concept maps for personalized learning path generation. In *Conference of the Italian Association for Artificial Intelligence*, pages 135–148. Springer, 2016. 69

[AL07] Marco Alfano and Biagio Lenzitti. Searching the internet for learning materials through didactic indicators. page 72, 06 2007. 15

[BP03] Peter Brusilovsky and Christoph Peylo. Adaptive and intelligent web-based educational systems. *Int. J. Artif. Intell. Ed.*, 13(2-4):159–172, April 2003. 17

[Bru01] P. Brusilowsky. User modeling and user-adapted interaction. *Adaptive hypermedia*, 11:87–110, 2001. 17

[BTM14] ALI REZA BANESHI, MAHNAZ DEHGHAN TEZERJANI, and HASAN MOKHTARPOUR. Grasha-richmann college students learning styles of classroom participation: role of gender and major. *Journal of advances in medical education & professionalism*, 2(3):103, 2014. 13

[CDG+08] Fay Chang, Jeffrey Dean, Sanjay Ghemawat, Wilson C. Hsieh, Deborah A. Wallach, Mike Burrows, Tushar Chandra, Andrew Fikes, and Robert E. Gruber. Bigtable: A distributed storage system for structured data. *ACM Trans. Comput. Syst.*, 26(2):4:1–4:26, June 2008. 18

[CJB99] Balakrishnan Chandrasekaran, John R Josephson, and V Richard Benjamins. What are ontologies, and why do we need them? *IEEE Intelligent Systems and their applications*, 14(1):20–26, 1999. 2

[DBp] DBpedia. Dbpedia. Last visited on 31 August 2016. 33, 69

[DJ13] V. Sree Dharinya and M.K. Jayanthi. Effective retrieval of text and media learning objects using automatic annotation. *World Applied Sciences Journal*, 27(1):123 – 129, 2013. 15

[DK12] Nada Dabbagh and Anastasia Kitsantas. Personal learning environments, social media, and self-regulated learning: A natural formula for connecting formal and informal learning. *The Internet and Higher Education*, 15(1):3 – 8, 2012. 16

[DMGL+16a] Carlo De Medio, Fabio Gasparetti, Carla Limongelli, Filippo Sciarrone, and Marco Temperini. Automatic extraction of prerequisites among learning objects using wikipedia-based content analysis. In *International Conference on Intelligent Tutoring Systems*, pages 375–381. Springer, 2016. 21

[DMGL+16b] Carlo De Medio, Fabio Gasparetti, Carla Limongelli, Filippo Sciarrone, and Marco Temperini. Mining prerequisite relationships among learning objects. In *International Conference on Human-Computer Interaction*, pages 221–225. Springer, 2016. 21

[DMGL+17a] Carlo De Medio, Fabio Gasparetti, Carla Limongelli, Matteo Lombardi, Alessandro Marani, Filippo Sciarrone, and Marco Temperini. Towards a characterization of educational material: An analysis of coursera resources. In Ting-Ting Wu, Rosella Gennari, Yueh-Min Huang, Haoran Xie, and Yiwei Cao, editors, *Emerging Technologies for Education*, pages 547–557, Cham, 2017. Springer International Publishing. 69

[DMGL+17b] Carlo De Medio, Fabio Gasparetti, Carla Limongelli, Filippo Sciarrone, and Marco Temperini. Course-driven teacher modeling for learning objects recommendation in the moodle lms. In *Adjunct Publication of the 25th Conference on User Modeling, Adaptation and Personalization*, pages 141–145. ACM, 2017. 53

[DMGL+18] Carlo De Medio, Fabio Gasparetti, Carla Limongelli, Filippo Sciarrone, Ilaria Torre, Giovanni Adorni, and Frosina Koceva. Wiki course

builder, a system for managing and sharing didactic material and concept maps. In *Proceedings of EDEN RW10 2018*, 2018. 69

[ECLLM16] Vladimir Estivill-Castro, Carla Limongelli, Matteo Lombardi, and Alessandro Marani. Dajee: A dataset of joint educational entities for information retrieval in technology enhanced learning. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 681–684. ACM, 2016. 20

[For10] Mary Forehand. Blooms taxonomy. *Emerging perspectives on learning, teaching, and technology*, 41:47, 2010. 53

[Gar05] Howard Gardner. *Educazione e sviluppo della mente. Intelligenze multiple e apprendimento*. Edizioni Erickson, 2005. 12

[GGP08] Ana Gil and Francisco Garca-Pealvo. Multiagent system for learning objects retrieval with context attributes. 33:320–326, 01 2008. 14

[GGPO15] Cornelia Gyrdi, Robert Gyorodi, George Pecherle, and Andrada Olah. A comparative study: Mongodb vs. mysql. 06 2015. 18

[Gil05] Jim Giles. Internet encyclopaedias go head to head, 2005. 2, 53

[GLS15] Fabio Gasparetti, Carla Limongelli, and Filippo Sciarrone. Wiki course builder: A system for retrieving and sequencing didactic materials from wikipedia. In *Information Technology Based Higher Education and Training (ITHET), 2015 International Conference on*, pages 1–6, June 2015. 3, 35

[GM09] Evgeniy Gabrilovich and Shaul Markovitch. Wikipedia-based semantic interpretation for natural language processing. *J. Artif. Int. Res.*, 34(1):443–498, March 2009. 17

[GML+18] Fabio Gasparetti, Carlo De Medio, Carla Limongelli, Filippo Sciarrone, and Marco Temperini. Prerequisites between learning objects: Automatic extraction based on a machine learning approach. *Telematics and Informatics*, 35(3):595–610, 2018. 22, 31

[GMS09] F. Gasparetti, A. Micarelli, and F. Sciarrone. A web-based training system for business letter writing. *Knowledge-Based Systems*, 22(4):287–291, May 2009. 19

[Gra94] Anthony F. Grasha. A matter of style: The teacher as expert, formal authority, personal model, facilitator, and delegator. *College Teaching*, 42(4):142–149, 1994. 40

[JB09] Kirsten A. Johnson and Jamie Bartolino. Creating community through the use of a class wiki. In *Proceedings of the 3D International Conference on Online Communities and Social Computing: Held As Part of HCI International 2009*, OCSC '09, pages 471–478, Berlin, Heidelberg, 2009. Springer-Verlag. 16

[JP13] Sonal Jain and Jyoti Pareek. Automatic extraction of prerequisites and learning outcome from learning material. *International Journal of Metadata, Semantics and Ontologies*, 8(2):145–154, 2013. 16

[KK05] Alice Y Kolb and David A Kolb. Learning styles and learning spaces: Enhancing experiential learning in higher education. *Academy of management learning & education*, 4(2):193–212, 2005. 12

[KM04] Rob Koper and Jocelyn Manderveld. Educational modelling language: modelling reusable, interoperable, rich and personalised units of learning. *British journal of educational technology*, 35(5):537–551, 2004. 69

[KRDK13] Judy Kay, Peter Reimann, Elliot Diebold, and Bob Kummerfeld. Moocs: So many learners, so much potential ... *IEEE Intelligent Systems*, 28:70–77, 2013. 1

[LDGS+13] Pasquale Lops, Marco De Gemmis, Giovanni Semeraro, Cataldo Musto, and Fedelucio Narducci. Content-based and collaborative techniques for tag recommendation: an empirical evaluation. *Journal of Intelligent Information Systems*, 40(1):41–61, 2013. 4

[LLM+15] C. Limongelli, M. Lombardi, A. Marani, F. Sciarrone, and M. Temperini. A recommendation module to help teachers build courses through the moodle learning management system. *New Review of Hypermedia and Multimedia*, 2(1-2):58–82, 2015. 53

[LLM+16] Carla Limongelli, Matteo Lombardi, Alessandro Marani, Filippo Sciarrone, and Marco Temperini. Concept maps similarity measures for educational applications. In *International Conference on Intelligent Tutoring Systems*, pages 361–367. Springer, 2016. 69

[LM10] Avinash Lakshman and Prashant Malik. Cassandra: A decentralized structured storage system. *SIGOPS Oper. Syst. Rev.*, 44(2):35–40, April 2010. 18

[LMST12] Carla Limongelli, Alfonso Miola, Filippo Sciarrone, and Marco Temperini. Supporting teachers to retrieve and select learning objects for personalized courses in the moodle ls environment. In *ICALT-2012: The 14th International Conference on Advanced Learning Technologies*, pages 518–520, Los Alamitos, CA, USA, 2012. IEEE Computer Society. 17

[LSV11] Carla Limongelli, Filippo Sciarrone, and Giulia Vaste. Personalized e-learning in moodle: the moodle ls system. *Journal of e-Learning and Knowledge Society*, 7(1):49–58, 2011. 14

[LWHG15] Chen Liang, Zhaohui Wu, Wenyi Huang, and C. Lee Giles. Measuring prerequisite relations among concepts. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1668–1674, Lisbon, Portugal, September 2015. Association for Computational Linguistics. 16, 28, 31

[MGL+16] Carlo De Medio, Fabio Gasparetti, Carla Limongelli, Matteo Lombardi, Alessandro Marani, Filippo Sciarrone, and Marco Temperini. Discovering prerequisite relationships among learning objects: A coursera-driven approach. In *ICWL*, 2016. 69

[MGLS17] C. De Medio, F. Gasparetti, C. Limongelli, and F. Sciarrone. Automatic extraction and sequencing of wikipedia pages for smart course building. In *2017 21st International Conference Information Visualisation (IV)*, pages 378–383, July 2017. 21

[Mit97] Thomas M. Mitchell. *Machine Learning*. McGraw-Hill, Inc., New York, NY, USA, 1 edition, 1997. 27

[MLR+18] Liyang Mao, Ou Lydia Liu, Katrina Roohr, Vinetha Belur, Matthew Mulholland, Hee-Sun Lee, and Amy Pallant. Validation of automated scoring for a formative assessment that employs scientific argumentation. *Educational Assessment*, 23(2):121–138, 2018. 69

[MMK+13] Stephen Maloney, Alan Moss, Jennifer Keating, George Kotsanas, and Prue Morgan. Sharing teaching and learning resources: perceptions of

a university's faculty members. *Medical education*, 47(8):811–819, 2013. 1

[MR14] S. Miranda and P. Ritrovato. Automatic extraction of metadata from learning objects. In *Intelligent Networking and Collaborative Systems (INCoS), 2014 International Conference on*, pages 704–709, Sept 2014. 15

[MR15] Sergio Miranda and Pierluigi Ritrovato. Supporting learning object repository by automatic extraction of metadata. *Journal of e-Learning and Knowledge Society*, 11(1), January 2015. 15

[MW13] David Milne and Ian H. Witten. An open-source toolkit for mining wikipedia. *Artif. Intell.*, 194:222–239, jan 2013. 17

[PM09] Portia Pusey and Gabriele Meiselwitz. Heuristics for implementation of wiki technology in higher education learning. In *Proceedings of the 3D International Conference on Online Communities and Social Computing: Held As Part of HCI International 2009*, OCSC '09, pages 507–514, Berlin, Heidelberg, 2009. Springer-Verlag. 16

[PM11] Portia Pusey and Gabriele Meiselwitz. Assessments in large- and small-scale wiki collaborative learning environments: Recommendations for educators and wiki designers. In *Proceedings of the 4th International Conference on Online Communities and Social Computing*, OCSC'11, pages 60–68, Berlin, Heidelberg, 2011. Springer-Verlag. 16

[Poc17] A. Poce. *Verba sequentur. Pensiero e scrittura per uno sviluppo critico delle competenze nella scuola secondaria*. Ricerche sperimentali. Franco Angeli, 2017. 68

[Ram03] Juan Enrique Ramos. Using tf-idf to determine word relevance in document queries. 2003. 38

[RC12] J. Russell and R. Cohn. *Hypertable*. Book on Demand, 2012. 18

[Rei06] Silvan Reinhold. Wikitrails: Augmenting wiki structure for collaborative, interdisciplinary learning. In *Proceedings of the 2006 International Symposium on Wikis*, WikiSym '06, pages 47–58, New York, NY, USA, 2006. ACM. 16

[RLM07] Stephan Repp, Serge Linckels, and Christoph Meinel. Towards to an automatic semantic annotation for multimedia learning objects. In *Proceedings of the International Workshop on Educational Multimedia and Multimedia Education*, Emme '07, pages 19–26. ACM, 2007. 15

[RSG08] Devshri Roy, Sudeshna Sarkar, and Sujoy Ghose. Automatic extraction of pedagogic metadata from learning content. *Int. J. Artif. Intell. Ed.*, 18(2):97–118, April 2008. 4, 15

[SFD11] Boutheina Smine, Rim Faiz, and Jean-Pierre Desclés. A semantic annotation model for indexing and retrieving learning objects. *JDIM*, 9(4):159–166, 2011. 15

[Siv12] Swaminathan Sivasubramanian. Amazon dynamodb: A seamlessly scalable non-relational database service. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, SIGMOD '12, pages 729–730, New York, NY, USA, 2012. ACM. 18

[SMD+10] Kateřina Slaninová, Jan Martinovič, Pavla Dráždilová, Gamila Obadi, and Václav Snášel. Analysis of social networks extracted from log files. In *Handbook of Social Network Technologies and Applications*, pages 115–146. Springer, 2010. 41

[SP06] Michael Strube and Simone Paolo Ponzetto. Wikirelate! computing semantic relatedness using wikipedia. In *Proceedings of the 21st National Conference on Artificial Intelligence - Volume 2*, AAAI'06, pages 1419–1424. AAAI Press, 2006. 17

[SSG14] Richard Scheines, Elizabeth Silver, and Ilya Goldin. Discovering prerequisite relationships among knowledge components. In J. Stamper, Z. Pardos, M. Mavrikis, and B.M. McLaren, editors, *Proceedings of the 7th International Conference on Educational Data Mining*, pages 355–356. ELRA, May 2014. 15, 20

[ST13] A. Sterbini and M. Temperini. Openanswer, a framework to support teacher's management of open answers through peer assessment. pages 164–170. IEEE, 2013. 17

[TMCO15] Tommaso Turchi, Alessio Malizia, Paola Castellucci, and Kai Olsen. Collaborative information seeking with ant colony ranking in real-time. In *11th Italian Research Conference on Digital Libraries - IRCDL 2015 Bozen-Bolzano, Italy, 29-30 January, 2015*, 2015. 17

[VMO+12] Katrien Verbert, Nikos Manouselis, Xavier Ochoa, Martin Wolpers, Hendrik Drachsler, Ivana Bosnic, and Erik Duval. Context-aware recommender systems for learning: a survey and future challenges. *IEEE Transactions on Learning Technologies*, 5(4):318–335, 2012. 1

[WFHP16] Ian H Witten, Eibe Frank, Mark A Hall, and Christopher J Pal. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2016. 30

[WL16] Shuting Wang and Lei Liu. Prerequisite concept maps extraction for automaticassessment. In *Proceedings of the 25th International Conference Companion on World Wide Web*, pages 519–521. International World Wide Web Conferences Steering Committee, 2016. 16

[WPP+00] Daniel C West, J Richard Pomeroy, Jeanny K Park, Elise A Gerstenberger, and Jonathan Sandoval. Critical thinking in graduate medical education: a role for concept mapping assessment? *Jama*, 284(9):1105–1110, 2000. 69

[YLCM15] Yiming Yang, Hanxiao Liu, Jaime Carbonell, and Wanli Ma. Concept graph learning from educational data. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, pages 159–168. ACM, 2015. 16

[Zha10] Mingxin Zhang. Social network analysis: history, concepts, and research. In *Handbook of social network technologies and applications*, pages 3–21. Springer, 2010. 41

[ZLW+17] Mengxiao Zhu, Hee-Sun Lee, Ting Wang, Ou Lydia Liu, Vinetha Belur, and Amy Pallant. Investigating the impact of automated feedback on students scientific argumentation. *International Journal of Science Education*, 39(12):1648–1668, 2017. 69